Original Article

QJEP

Quarterly Journal of Experimental Psychology
2025, Vol. 78(4) 707–730

DOI: 10.1177/17470218241255916
qjep.sagepub.com
Sage

Revisiting representativeness heuristic classic paradigms: Replication and extensions of nine experiments in Kahneman and Tversky (1972)

Lewend Mayiwar[1]*, Kai Hin Wan[2]*, Erik Løhre[1] and Gilad Feldman[2]

Abstract

Kahneman and Tversky showed that when people make probability judgements, they tend to ignore relevant statistical information (e.g., sample size) and instead rely on a representativeness heuristic, whereby subjective probabilities are influenced by the degree to which a target is perceived as similar to (representative of) a typical example of the relevant population, class or category. Their article has become a cornerstone in many lines of research and has been used to account for various biases in judgement and decision-making. Despite the impact this article has had on theory and practice, there have been no direct replications. In a pre-registered experiment ($N = 623$; Amazon MTurk on CloudResearch), we conducted a replication and extensions of nine problems from Kahneman and Tversky's 1972 article. We successfully replicated eight out of the nine problems. We extended the replication by examining the consistency of heuristic responses across problems and by examining decision style as a predictor of participants' use of the representativeness heuristic. Materials, data, and code are available on: https://osf.io/nhqc4/

Keywords

representativeness heuristic; judgement and decision-making; subjective probability; replication; decision style

Received: 26 October 2023; revised: 5 March 2024; accepted: 2 April 2024

Kahneman and Tversky's heuristics and biases programme is considered a cornerstone of the fields of judgement and decision-making and behavioural economics and has had a profound impact on our understanding of human reasoning and on psychological research at large (Keren & Teigen, 2004; Thaler, 2016). Their work has also been highly influential in many applied contexts, including decision-making in medicine (Whelehan et al., 2020), project management (McCray et al., 2002), and entrepreneurship (Busenitz & Barney, 1997).

Kahneman and Tversky identified various heuristics, or mental shortcuts, that people rely on to make judgements and decisions, which usually work quite well, but might also lead to predictable and systematic deviations from the classic economics rationality model (Tversky & Kahneman, 1974). In a foundational article, Kahneman and Tversky (1972; hereafter referred to as KT) demonstrated that people often ignore statistical information (e.g., sample size) when judging the probability of events and instead rely on a *representativeness heuristic*: They judge the probability of an event based on how similar it is to a prototype or a stereotype. For example, when estimating the probability that an entrepreneur will be successful, the similarity of the entrepreneur to prototypically successful exemplars (energetic, confident, extraverted) may

[1]Department of Leadership and Organizational Behavior, BI Norwegian Business School, Oslo, Norway
[2]Department of Psychology, The University of Hong Kong, Pokfulam, Hong Kong SAR

*Shared first coauthor

Corresponding author:
Gilad Feldman, Department of Psychology, The University of Hong Kong, Pokfulam, Hong Kong 999077 SAR.
Email: gfeldman@hku.hk

receive more weight than objective information about the low base rate of success.

More formally, Kahneman and Tversky (1972) originally defined the use of the representativeness heuristic as evaluating "the probability of an uncertain event, or a sample, by the degree to which it is: (i) similar in essential properties to its parent population; and (ii) reflects the salient features of the process by which it is generated" (p. 431). In later work, Tversky and Kahneman (1982) made the definition more general by stating that "representativeness is a relation between a process or a model, M, and some instance or event, X, associated with that model" (p. 85), but for the purposes of this article, we stick to the original definition and the idea that subjective probability judgements can be highly influenced by the perceived similarity between the events/samples and populations/generation processes under consideration.

The target article from 1972 has been extremely influential with 8,717 Google Scholar citations as of February 2024. The representativeness heuristic has been used to explain a range of biases and errors in judgement and decision-making, such as the conjunction fallacy (Tversky and Kahneman, 1983), the hot hand and the gambler's fallacy (Sundali & Croson, 2006), and in many different domains, such as health care (Brannon & Carson, 2003), politics (Stolwijk & Vis, 2021), and financial markets (Fuster et al., 2010). The use and development of the representativeness heuristic have even been studied among preschool-aged children (Gualtieri & Denison, 2018).

Although there has been extensive empirical research on the representativeness heuristic, critics argue that the concept is vaguely defined and poorly understood (Galavotti et al., 2021; Gigerenzer, 1996). Furthermore, despite the wide use of the representativeness heuristic in different contexts, there have been few attempts to conduct direct independent replications of the original findings. Some studies (e.g., Bar-Hillel, 1984; Olson, 1976) have replicated some of the problems used in the target article, but to our knowledge, this is the first comprehensive replication of an article that covers a wide range of different empirical approaches to the same underlying phenomenon.

Following the growing recognition of the importance of replications in psychological research (e.g., Nosek et al., 2022; Zwaan et al., 2018), we aimed to revisit and reassess the robustness of Kahneman and Tversky's foundational study by conducting an independent well-powered preregistered replication with extensions.

We were also motivated by the potential for methodological improvements to the target article. The article by Kahneman and Tversky (1972) was published over 50 years ago. Therefore, they did not report "new statistics" (such as effect sizes) to allow for easier follow-up research and applications, and they presented several null hypotheses that were tested using null hypothesis significance testing which was not meant to quantify the null. Our replication improves on those points.

We also aimed for a more robust understanding of the phenomenon. The target article reported multiple studies focusing on different problems which were completed by different samples. We had participants complete multiple problems, thus allowing us to also examine people's consistency in responses across multiple problems and begin to map associations between them. In the heuristics and biases literature, problems are almost exclusively studied between subjects, with only a few studies investigating the coherence of different heuristics and biases (Ceschi et al., 2019). However, very little research has investigated whether different conceptualizations of the same proposed underlying heuristic show internal consistency. Indeed, we know very little about the internal consistency of JDM tasks (Parsons et al., 2019). Recent studies found low internal consistency of different measures of heuristic responses like anchoring (e.g., Röseler et al., 2022). This project is part of systematic replications of seminal review articles examining many paradigms of a broad phenomenon, such as of mental accounting in Thaler (1999) (Li & Feldman, 2022), of "goals as reference points" in Heath et al. (1999) (Au & Feldman, 2020), the "belief in the law of small numbers" in Tversky and Kahneman (1971) (Hong & Feldman, 2023), and another seminal article on the representativeness heuristic by Kahneman and Tversky (1972) (Chan & Feldman, 2024). The systematic replication of studies reported on a single phenomenon resulted in valuable insights mapping differences in strength across different methods and contexts and assessing overall consistency.

Finally, several studies have sought to identify contextual factors that may predict reliance on the representativeness heuristic (e.g., Agnoli, 1991; Cox & Mouw, 1992; Grether, 1992). However, less is known about how personality-level predictors relate to the use of heuristics like representativeness. We extended the replication by also including decision style (i.e., preference for intuitive and analytical thinking) as a potential predictor of participants' use of the representativeness heuristic. According to Kahneman and Tversky (1972), the reliance on representativeness is a type of heuristic, or an intuitive response (Kahneman & Frederick, 2002). Kahneman and Tversky (1972) suggested that people's intuitive probability judgements often do not follow laws of probability and chance, as these are not incorporated in our intuitive thinking.

The intuitive decision style has been characterised by a reliance on quick and effortless thinking based on hunches and feelings (Harren, 1979). On the other hand, the analytical decision style is characterised by careful and deliberate search for a logical evaluation of alternatives (Harren, 1979). In contrast to the intuitive style, the analytical style has been associated with lower susceptibility to various decision biases (e.g., Chatterjee et al., 2000; Smith & Levin, 1996) and greater performance in a range of different tasks (Alaybek et al., 2022). We, therefore, hypothesised that participants' reliance on the representativeness heuristic would be positively predicted by the intuitive decision style and negatively predicted by the analytical style.

Transparency statement

We provided all materials, data, and code at: https://osf.io/nhqc4/. All measures, manipulations, and exclusions conducted for this investigation are reported.

We pre-registered the study, which can be accessed at: https://osf.io/57rmd/. We wrote the pre-registration for this project in a "Registered Report" format, using the Registered Report template by Feldman (2023), meaning that the pre-registration was a manuscript with a results section written with a simulated random dataset, analysis code, and included an exported Qualtrics survey (see https://osf.io/57rmd/files/osfstorage for files, https://osf.io/nbdjr for the main manuscript, https://osf.io/k2tqy for the Qualtrics survey, and https://osf.io/g4nja for the planned analysis code and simulated datasets). Deviations from the pre-registration are listed in sub-section "Deviations from pre-registration" of the Method section. We did not perform any analyses before completing the data collection.

We analysed the data in *R* (version 4.3.2, R Core Team, 2023), with *haven* version 2.5.4 (Wickham et al., 2023), *tidyverse* version 2.0.0 (Wickham et al., 2019), *ggplot2* version 3.4.4 (Wickham, 2016), *psych* version 2.3.12 (Revelle, 2024), *emmeans* version 1.9.0 (Lenth et al., 2023), *BayesFactor* version 0.8.12-4.6 (Morey & Rouder, 2018), *cowplot* version 1.1.2 (Wilke, 2020), *ggpubr* version 0.6.0 (Kassambara, 2020), Superpower (Lakens & Caldwell, 2021), and *kableExtra* version 1.3.4.9 (Zhu, 2021). Effect size, power, and confidence intervals were all calculated with the help of a guide by Jané et al. (2024).

Method

Power analysis

We conducted a power analysis for each problem separately. Whenever possible, we computed effects from the information and descriptives reported in the target article, and the code is provided in the OSF with additional details in the online Supplementary Material (section "Power analysis of target article effects"). We used the smallest effect size out of all problems, and then halved it for a more conservative estimate. The smallest observed effect size was in Problem 4 (Cohen's $h = 0.39$), which we then halved, resulting in a Cohen's h of about 0.20. With 95% power and 5% alpha error rate, we needed a sample size of 334. This was the largest required sample size among all analyses.

We were worried that answering all nine problems would be too cognitively demanding on our participants, and so to reduce possible cognitive fatigue we assigned participants to answer only five randomly selected problems out of nine problems. We therefore doubled our target sample size estimate, resulting in a target sample size of 668 participants.

Participants

We used the CloudResearch platform (formerly known as TurkPrime; Litman et al., 2017) to recruit a total of 683 American Amazon Mechanical Turk (MTurk) workers. We pre-registered the following exclusion criteria: Participants indicating low proficiency in English (< 5, on a 1–7 scale), participants who reported not being serious about filling in the survey (< 4, on a 1–5 scale), participants who correctly guessed the hypothesis of this study in the funnelling section, participants who had already seen the material in the survey before, participants who failed to complete the survey, and participants who failed two attention checks that were embedded in the questionnaire ("one hundred is more than fifty" and "fifty is more than one hundred"). Results were almost identical when using the full sample, with a slight deviation in Problems 7–9 (Likelihood of Sampling Outcomes): In the analysis with exclusions, all effects had confidence intervals that included the SESOI, whereas in the analysis without exclusions, all but two effects had confidence intervals that included the SESOI (this was only the case when using an alternative non-pre-registered approach to calculate the SESOI).

We note that we initially pre-registered to focus our reporting on the full sample (pre-exclusion). However, we deviated from the pre-registered plan and instead chose to present post-exclusion results in the main text and pre-exclusion results in the online Supplementary Material. This decision was made to ensure the exclusion of participants who struggled to maintain concentration during the experiment, which is cognitively demanding, due to the numerous statistics-heavy tasks.

The sample following the pre-registered exclusion criteria included 623 participants ($M_{age} = 42.7$, $SD = 12.5$; 317 males, 303 females, three "other" or "would rather not disclose"). Given that participants only answered five out of the nine problems, it means that each problem had, on average, 346 participants. We conducted a sensitivity analysis for all tests except Problem 6 which used a Kolmogorov-Smirnov test. The sample size provided 95% power (5% alpha) to detect a Cohen's $h = 0.27$ for a one-sample proportion test (Problems 1, 2, 4, and 7–9), a Cohen's $d = 0.19$ for a paired-samples t-test (Problem 11), and a Cohen's $f = 0.26$ for an ANOVA with 10 conditions (Cohen's $f = 0.23$ in G*Power). The online Supplementary Material includes power curve plots. We summarised a comparison of the target article's sample and the replication's sample in Table 1.

Table 1. Sample comparison between the target article and the replication.

	Kahneman and Tversky (1972)	Replication sample
Sample size	Approximately 1,500 in total. Different participants responded to different problems, with some responding to 2–4 problems. Sample size in each problem: Problem 1 = 92 Problem 2 = 89 Problem 4 = 52 Problem 6 = 558 divided into 9 conditions Problems 7–9 = 97 divided two conditions Problem 10 = 560 divided into 10 conditions Problem 11 = 115	623 in total (randomised to five out of nine problems). Sample size in each problem: Problem 1 = 343 Problem 2 = 360 Problem 4 = 348 Problem 6 = 593 divided into 9 conditions Problem 7 = 346 Problem 8 = 346 Problem 9 = 346 divided into two conditions Problem 10 = 346 divided into 10 conditions Problem 11 = 313
Type of sample	High-school students (Problems 1–4 and 8) Undergraduates (Problems 5–7 and 11–12)	MTurk online workers on CloudResearch
Geographic origin	Israel (Problems 1–4 and 8) United States (Problem 5–7 and Problems 11–12)	United States
Gender	Not specified	317 males, 303 females, 3 other/would rather not disclose
Median age (years)	Not specified	40
Average age (years)	Not specified	42.7
Age range (years)	15–18 (Israeli high school students), not specified (other samples)	21–91
Medium (location)	Pen and paper in a classroom situation	Computer (online)
Compensation	Not specified	Nominal payment
Year	Not specified	2020

Procedure

The target article had multiple studies with both experimental manipulations and one-sample experiments.

In our replication, participants were randomly assigned to respond to a subset of five out of nine chosen problems from the target article. Participants first indicated their consent and their understanding and willingness to participate in the study. Next, they answered the Decision Styles Scale (DSS), and proceeded to complete the decision problems (summarised in Table S4 in the online Supplementary Material). Finally, participants provided demographic information and were debriefed.

Overview of problems

Kahneman and Tversky (1972) used 11 problems to test their hypotheses. They found support for all of the hypotheses in all problems. We included all problems except for Problem 3 because it required a specific group of participants and statistical knowledge (which was originally from Tversky & Kahneman's, 1971 "law of small numbers," and addressed in a separate replication in Hong & Feldman, 2023), and Problem 5 which was an anecdotal/illustrative example and did not report results. For the sake of simplicity and consistency, we assigned numerical labels to the problems in our article, while preserving the original order in which the problems were presented. We did not change the numbering of the problems despite not including Problems 3 and 5.

To ease reading, in the sections below we grouped the nine problems that were included in this replication into five groups based on domain: Problems 1 and 2 (sample-to-population similarity), Problem 4 (reflection of randomness), Problem 6 (sampling distributions), Problems 7–9 (likelihood of sampling outcomes), Problems 10 and 11 (posterior probabilities).

Problems' study design

Problems 1 and 2 (sample-to-population similarity), Problem 4 (reflection of randomness), and Problem 11 (posterior probabilities, non-binomial) were all one-sample experiments that involved no manipulations.

Some of the problems involved testing a null hypothesis. KT interpreted not meeting the significance threshold for differences between the conditions as support for the null hypothesis (which is why *p*-values were large and the effect sizes small in these problems). We complemented their approach with more appropriate methods that quantify the null, specifically using equivalence testing and Bayesian analyses (Lakens 2017). The target article did not report effect sizes, but we computed these in the problems that had sufficient statistical information (see accompanying RMarkdown code files, and section "Effect size calculations of the original study effects" in the online

Table 2. Summary of target article's findings.

Problem	Factors		*p*	Effect [95% CI]
1. Sample-population Similarity (Birth Sequence)	/		< .01	Cohen's h = 0.68 [0.48, 0.89]
2. Sample-population Similarity (High-school Prog.)	/		< .01	Cohen's h = 0.53 [0.32, 0.74]
4. Reflection of Randomness	/		.008	Cohen's h = 0.39 [0.12, 0.67]
6. Sampling Distributions	Gender distribution	N = 10 vs N = 100	1.00	$D = 0.09$
		N = 100 vs N = 1,000	.957	$D = 0.18$
		N = 10 vs N = 1,000	.990	$D = 0.18$
	Heartbeat distribution	N = 10 vs N = 100	.978	$D = 0.18$
		N = 100 vs N = 1,000	.957	$D = 0.27$
		N = 10 vs N = 1,000	.959	$D = 0.18$
	Height distribution	N = 10 vs N = 100	1.00	$D = 0.00$
		N = 100 vs N = 1,000	.944	$D = 0.29$
		N = 10 vs N = 1,000	.944	$D = 0.29$
7. Likelihood of Sampling Outcomes (Babies)	"More extreme" condition [a]		.968	Cohen's h =-0.30 [-0.58, -0.03]
	"Less extreme" condition		.959	Cohen's h =-0.30 [-0.60, -0.01]
8. Likelihood of Sampling Outcomes (Investigator)	"More extreme" condition		.103	Cohen's h = 0.20 [-0.08, 0.48]
	"Less extreme" condition		.954	Cohen's h =-0.28 [-0.57, 0.00]
9. Likelihood of Sampling Outcomes (Disease)	"More extreme" condition		.323	Cohen's h = 0.09 [-0.20, 0.37]
	"Less extreme" condition		.677	Cohen's h =-0.09 [-0.37, 0.19]
10. Posterior Probability (Binomial)	Initial proportion: 5:1	5:1 vs 4:2	< .01	
		5:1 vs 8:4	< .01	
		5:1 vs 40:20	< .01	
		18:14 vs 4:2	< .01	
		18:14 vs 8:4	< .01	
		18:14 vs 40:20	< .01	
	Initial proportion: 2:1	5:1 vs 4:2	< .01	
		5:1 vs 8:4	< .01	
		5:1 vs 40:20	< .01	
		18:14 vs 4:2	< .01	
		18:14 vs 8:4	< .01	
		8:14 vs 40:20	< .01	
11. Posterior Probability (Non-binomial)	/		< .01	

Note. D is the Kolmogorov-Smirnov test statistic. Problems 6, 7, 8, and 9 tested null hypotheses. Therefore, p-values were large and effect sizes were small, and reflect a one-tail t-test of the directionality of the prediction (which is why confidence intervals might not include the null, yet have very high p-values).

[a] More extreme condition = Outcome more extreme than the specified mean of probability, Less extreme condition = Outcome less extreme than the specified mean of probability.

Supplementary Material). We summarised the target article's findings in Table 2.

Problems

We summarised all the problems, their designs, predictions about participants' answers, and accurate answers in Table 3. We briefly describe those in the next subsections.

Sample-to-population similarity (problems 1 and 2)

Kahneman and Tversky designed two tasks to identify the characteristics of a sample that makes it representative of a population. No experimental manipulation was used in these problems.

Birth sequence (problem 1). The problem evaluated the characteristics of a sample that makes it representative (i.e., the similarity of the sample to the population). Participants were informed that all families with six children in a city had been surveyed and that 72 six-children families had the birth order of boys and girls of GBGBBG. The participants were then asked to estimate the number of families with the birth order of boys and girls BGBBBB.

Although both sequences are equally likely (each sequence represents a random arrangement of births, with each birth being independent of the others), KT

Table 3. Replication: problems, design, and predictions.

#	Domain	Design	Problem	Predictions	Correct answer
1.1	Similarity of sample to population (birth sequence)	No manipulation; one sample	All families of six children in a city were surveyed. In 72 families, the exact order of births of boys and girls was G B G B B G. What is your estimate of the number of families surveyed in which the exact order of births was B G B B B B / B B B G G G / G B B G B G?	1a: Sample with boy-girl split closer to expected equal 50–50 split in the population (GBGBBG) is perceived as more probable than a lesser equal split sequence (BGBBBB) 1b: Sample with less orderly sequence (GBBGBG) is perceived as more probable than a sample with an orderly sequence (BBBGGG)[a]	"The two birth sequences are about equally likely" (p. 432) 1a: equal probability 1b: equal probability.
2	Similarity of sample to population (gender proportion)	No manipulation; one sample	There are two programmes in a high school. Boys are a majority (65%) in programme A, and a minority (45%) in programme B. There is an equal number of classes in each of the two programmes. You enter a class at random, and observe that 55% of the students are boys. What is your best guess—does the class belong to programme A or to programme B?"	2: When observing a class with 55% boys, class is perceived to be more likely programme A (65% boys) than programme B (45% boys) given that boys are a majority and therefore more "representative".	"In fact, it is slightly more likely that the class belongs to program B (since the variance for $p = .45$ exceeds that for $p = .65$)." (p. 433)
4	Reflection of randomness in the sample	No manipulation; one sample	On each round of a game, 20 marbles are distributed at random among five children: Alan, Ben, Carl, Dan, and Ed. Consider the following distributions. Type I Type II Alan 4 4 Ben 4 4 Carl 5 4 Dan 4 4 Ed 3 4 In many rounds of the game, will there be more results of type I or of type II?	Type II distribution is perceived as more probable than Type II distribution.	"The uniform distribution of marbles (II) is, objectively, more probable than the nonuniform distribution (I)" (p. 434)
6a	Sampling Distributions: Distribution of Sexes (Binomial, $p = .50$)	3 conditions between-subject (sample size): $N = 10$, $N = 100$, $N = 1,000$	[10/100/1,000] babies are born everyday in a certain region. Given that the possibilities of both gender are equal (50/50), on what percentage of days will the number of boys among [10/100/1,000] babies be as follows: (Note that the categories include all possibilities, so your answers should add up to about 100%). __ [0 boys/Up to 5 boys/Up to 50 boys] (1) __ [1 boy/5 to 15 boys/50 to 150 boys] (2) __ (2 boys/15 to 25 boys/150 to 250 boys [3] __ (3 boys/25 to 35 boys/250 to 350 boys [4] __ (4 boys/35 to 45 boys/350 to 450 boys [5] __ (5 boys/45 to 55 boys/450 to 550 boys [6] __ (6 boys/55 to 65 boys/550 to 650 boys [7] __ (7 boys/65 to 75 boys/650 to 750 boys [8] __ (8 boys/75 to 85 boys/750 to 850 boys [9] __ (9 boys/85 to 95 boys/850 to 950 boys [10] __ (10 boys/More than 95 boys/More than 950 boys [11] Note: The means of estimate of each row of each subject were taken to make the mean sampling distributions.	[KT's null effect hypothesis] Law of small numbers / Sample size neglect: There would be no differences in distribution comparing condition with 10, 100, or 1,000. [Competing, reframed from the null effect] Law of big numbers / Sample size sensitivity There would be differences in distribution comparing condition with 10, 100, or 1,000.	

(Continued)

Table 3. (Continued)

#	Domain	Design	Problem	Predictions	Correct answer
6b	Sampling Distributions: Distribution of Heartbeat Type (Binomial, $p = .80$)	3 conditions between-subject (sample size): $N = 10$, $N = 100$, $N = 1{,}000$	[10/100/1,000] babies are born everyday in a certain region. Given that 80% of all newborns have a heartbeat of type α and the remaining 20% have a heartbeat of type β, on what percentage of days will the number of babies with heartbeat of type α among [10/100/1,000] be as follows (Note that the categories include all possibilities, so your answers should add up to about 100%). __ [0 babies/Up to 5 babies/Up to 50 babies] (1) __ [1 baby/5 to 15 babies/50 to 150 babies] (2) __ (2 babies/15 to 25 babies/150 to 250 babies [3] __ (3 babies/25 to 35 babies/250 to 350 babies [4] __ (4 babies/35 to 45 babies/350 to 450 babies [5] __ (5 babies/45 to 55 babies/450 to 550 babies [6] __ (6 babies/55 to 65 babies/550 to 650 babies [7] __ (7 babies/65 to 75 babies/650 to 750 babies [8] __ (8 babies/75 to 85 babies/750 to 850 babies [9] __ (9 babies/85 to 95 babies/850 to 950 babies [10] __ (10 babies/More than 95 babies/More than 950 babies [11] Note: The means of estimate of each row of each subject were taken to make the mean sampling distributions.	[KT's null effect hypothesis] Law of small numbers / Sample size neglect: There would be no differences in distribution comparing condition with 10, 100, or 1,000. [Competing, reframed from the null effect] Law of big numbers / Sample size sensitivity There would be differences in distribution comparing condition with 10, 100, or 1,000.	
6c	Sampling Distributions: Distribution of height.	3 conditions between-subject (sample size): $N = 10$, $N = 100$, $N = 1{,}000$	A regional induction centre records the average height of the [10/100/1,000] men who are examined every day. Given that the average height of the male population lies between 170–175 cm and the frequency of heights decreases with the distance from the mean, on what percentage of men's different height classes will be recorded on a certain day as follows: __ Up to 160 cm (1) __ 160–165 cm (2) __ 165–170 cm (3) __ 170–175 cm (4) __ 175–180 cm (5) __ 180–185 cm (6) __ More than 185 cm (7((Note that the categories include all possibilities, so your answers should add up to about 100%)	[KT's null effect hypothesis] Law of small numbers / Sample size neglect: There would be no differences in distribution comparing condition with 10, 100, or 1,000. [Competing, reframed from the null effect] Law of big numbers / Sample size sensitivity There would be differences in distribution comparing condition with 10, 100, or 1,000.	
7	Likelihood of Sampling Outcomes in Small vs. Large Samples Size of hospital	2 conditions between-subject (more versus less)	A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower. For a period of 1 year, each hospital recorded the days on which [more/less] than 60% of the babies born were boys. Which hospital do you think recorded more such days? (The larger hospital/The smaller hospital/About the same (i.e., within 5% of each other.	People tend to judge the two hospitals as having the same likelihood for 60% boys.	Smaller hospital has larger variance and therefore more likely to have a day with 60%.
8	Likelihood of Sampling Outcomes in Small vs. Large Samples Line vs. page	2 conditions between-subject (more versus less)	An investigator studying some properties of language selected a paperback and computed the average word-length in every page of the book (i.e., the number of letters in that page divided by the number of words). Another investigator took the first line in each page and computed the line's average word-length. The average word-length in the entire book is 4. However, not every line or page has exactly that average. Some may have a higher average word-length, some lower. The first investigator counted the number of pages that had an average word-length of 6 or [more/less] and the second investigator counted the number of lines that had an average word-length of 6 or [more/less]. Which investigator do you think recorded a larger number of such units (pages for one, lines for the other)? (The page investigator; The line investigator; About the same [i.e., within 5% of each other])	People tend to judge the two investigators as having the same likelihood of having an average of 6 or more words per unit.	Line has smaller sample and larger variance and therefore more likely to have average word-length of 6 or more than page.

(Continued)

Table 3. (Continued)

#	Domain	Design	Problem	Predictions	Correct answer
9	Likelihood of Sampling Outcomes in Small vs. Large Samples 3 men versus 1 man	2 conditions between-subject (more versus less)	A medical survey is being held to study some factors pertaining to coronary diseases. Two teams are collecting data. One checks 3 men a day, and the other checks 1 man a day. These men are chosen randomly from the population. Each man's height is measured during the checkup. The average height of adult males is 5 ft 10 in., and there are as many men whose weight is above average as there are men whose height is below average. The team checking 3 men a day ranks them with respect to their height, and count the days on which the height of the middle man is [more/less] than 5 ft 11 in. The other team checking 1 man a day merely counts the days on which the man they checked was [taller/shorter] than 5 ft 11 in. Which team do you think counted more such days? The team checking 3 men; The team checking 1 man; About the same (i.e., within 5% of each other)	People tend to judge the medical surveys as having the same likelihood of men taller than 5 ft 10 in.	1 man a day is smaller and has larger variance than 3 men a day, and therefore more likely to record more days where the height is taller than 5 ft 11 in.
10	Posterior Probabilities Binomial task	2 x 5 between-participants design	Consider two very large decks of cards, denoted A and B. In deck A, [5/6; 2/3; 5/6; 2/3; 5/6; 2/3; 5/6; 2/3; 5/6; 2/3] of the cards are marked X and [1/6; 1/3; 1/6; 1/3; 1/6; 1/3; 1/6; 1/3; 1/6; 1/3] are marked O. In deck B, [1/6; 1/3; 1/6; 1/3; 1/6; 1/3; 1/6; 1/3; 1/6; 1/3] of the cards are marked X, and [5/6; 2/3; 5/6; 2/3; 5/6; 2/3; 5/6; 2/3; 5/6; 2/3] are marked O. One of the decks has been selected by chance, and [12; 12; 6; 6; 60; 60; 6; 6; 32; 32] cards have been drawn at random from it, of which [8; 8; 4; 4; 40; 40; 5; 5; 18; 18] are marked X and [4; 4; 2; 2; 20; 20; 1; 1; 14; 14] are marked O. What do you think the probability is that the [12; 12; 6; 6; 60; 60; 6; 6; 32; 32] cards were drawn from deck A, that is from the deck in which most of the cards are marked X? For example, if you think that there is a 100% chance that the sample was drawn from deck A, you can input "1". If you think that there is a 60% chance that the sample was drawn from deck A, you can input "0.6".	People tend to rely on sample proportions of the two objects (as this is the most representative feature). In both pairs of population proportions (5/6 and 1/6 vs. 2/3 and 1/3), participants' posterior estimates in the 5:1 sample proportion condition would be larger than in the 4:2, 8:4, and 40:20 conditions, which would be larger than in the 18:14 conditions.	"In the symmetric binomial task the objective posterior probability depends only on the difference between the numbers of red and blue chips observed in the sample. posterior odds are given by (p/l-p)^(r-b)" (p. 446–8)
11	Posterior Probabilities Non-Binomial Task	2 conditions within-participants (single person vs. 6 persons)	The average heights of adult males and females in the US are, respectively, 5 ft 10 in. and 5 ft 4 in. Both distributions are approximately normal with a standard deviation of about 2.5 in. An investigator has selected one population by chance and has drawn from it a random sample. (i) What do you think is the probability in percentage that he has selected the male population if the sample consists of a single person whose height is 5 ft 10 in.? (ii) What do you think is the probability in percentage that he has selected the male population if the sample consists of 6 persons whose average height is 5 ft 8 in.?	In a population with average heights of 5 ft 10 in. for males and 5 ft 4 for females, people tend to perceive a randomly drawn single person with 5 ft 10 in. as more likely to drawn from a male population than randomly drawn 6 persons averaging 5 ft 8 in.	"The correct odds are 16% in case (i) and 29% in case (ii)." (p. 449)

[a] The predictions here do not compare the estimated number of families with sequence BGBBBB against the other two sequences listed under the Problem column (under "What is your estimate. . ."). Instead, it presents two distinct predictions made by KT: one comparing BGBBBB against the stated number of surveyed families with the sequence BBBGGG, and the other comparing the sequences BBBGGG and GBBGBG against each other.

hypothesised that the sequence BGBBBB would be judged as less probable than GBGBBG. This is because GBGBBG exhibits an equal distribution of boys and girls (half boys and half girls), which is more representative of the population.

In addition, the target article added two questions: one estimating the frequency of BBBGGG and the other estimating the frequency of GBBGBG. KT included the two questions to investigate whether people ignore order information and base their judgements only on the frequency of boys and girls. Specifically, they hypothesised that participants would judge the sequence BBBGGG as less likely than GBBGBG because BBBGGG looks too organised, even though both sequences have the same frequency of boys and girls.

High-school programme (problem 2). The problem presented participants with two programmes in a school. 65% of the population of programme A are boys and 45% of the population of programme B are boys. Participants were asked to determine if a class with 55% of boys is more likely to belong to programme A or programme B.

KT hypothesised that more participants would think the class belongs to programme A because boys represent the majority of students in both the class and programme A, and thus the class maintains the majority/minority relation of the population. However, the class has a higher probability of belonging to programme B than programme A as the variance of programme B (p=.45) is larger than the one of programme A (p=.65). Programme B has greater variability because its proportion of boys (45%) is closer to the midpoint (50%) of all possibilities, allowing for more potential fluctuation in the gender composition. In contrast, programme A, with its majority of boys (65%), offers less room for variation as any change would likely maintain a majority of boys.

Reflection of randomness (problem 4)

Participants were shown two distributions of 20 marbles that were randomly distributed to five kids. In Distribution "Type I," three children received four marbles, one child received five marbles, and one child received three marbles. In Distribution "Type II," all five children received four marbles. Participants were asked to select the distribution that was more probable. KT hypothesised that people would judge the "Type I" Distribution as more likely, although both distributions are statistically equally random, or equally likely. If anything, Distribution "Type II" is more likely because all children receive the same number of marbles, making it more probable overall. Distribution "Type I" appears more random because it does not follow a clear pattern. But this does not mean it is statistically more random—the "randomness" one recognises is not necessarily the same as true randomness in statistics. When there are regularities or clusters, we tend to perceive them as non-random.

Sampling distributions (problem 6)

Problem 6 involved a between-subjects design with three distinct scenarios: gender distribution (binomial with probability 0.5), heartbeat types (binomial with probability 0.8), and average height (non-binomial distribution). Note that because participants were randomly assigned to complete five problems from the total pool, which included these three Problem 6 scenarios, each participant could be assigned one, two, or all three Problem 6 scenarios within their set of five problems. For each scenario, participants were assigned to one of three sample sizes (N = 10, N = 100, N = 1,000).

KT proposed that participants neglect sample size when determining the sampling distribution as sample size is not a salient feature of the population and, therefore, does not affect representativeness. Thus, KT hypothesised that across all scenarios and types of distribution, there would be no differences between the three sampling distributions (N=10, N=100, N=1,000), with probabilities of different outcomes judged according to their similarity to the population mean or proportion. For instance, KT explained that participants might judge the event of finding more than 600 boys in a sample of 1,000 babies as equally representative as finding more than 60 boys in a sample of 100 babies, even though the latter is much more likely.

The Problem 6 scenarios were designed to the absence of an effect, or a null hypothesis. The target article did not conduct any statistical tests for this problem and instead relied on descriptive results and a visual inspection of the distributions. In addition, the target article did not specify whether participants were randomly assigned to the three conditions. We evaluated the results based on the p-values and comparisons between the graphs in this replication and the target article.

Distribution of sexes (binomial, p *=.50; problem 6a).* Participants were told that [10/100/1,000] babies are born every day in a certain region. For instance, for N=1,000, the question read as follows: "On what percentage of days will the number of boys among 1,000 babies be as follows: 1) Up to 50 boys, 2) 50 to 150 boys, 3) 150 to 250 boys, [. . .] 10) 850 to 950 boys, 11) More than 950 boys; Note that the categories include all possibilities, so your answers should add up to about 100%.." For N=100, the 11 categories were: 1) up to 5, 2) 5–15, etc. For N=10, each category contained a single outcome, for instance, 1) 0 boys, 2) 1 boy, 3) 2 boys, etc. We note a weakness in the target's design that the categories were overlapping (e.g., 150 boys belonged to both 2 and 3) and so it is unclear how participants used these categories, yet we decided to follow the target's design and did not make any adjustments because we were not sure how such a change might impact results.

Distribution of heartbeat type (binomial, $p = .80$*; problem 6b).* Participants were told that [10/100/1,000] babies are born every day in a certain region and that 80% of all newborns have a heartbeat of type α and the remaining 20% have a heartbeat of type β. For each sample size, participants produced sampling distributions for the number of babies born every day with a heartbeat of type α using the same 11 categories as above only changed to refer to babies instead of boys.

Distribution of height (problem 6c). Participants were told that a regional induction centre records the average height of the [10/100/1,000] men who are examined every day. They were also told that the average height of the male population lies between 170 and 175 cm and that the frequency of heights decreases with the distance from the mean. For each sample size, participants produced a sampling distribution of average height in the following seven categories: up to 160 cm, 160–165 cm, . . ., and more than 185 cm. We note that we opted to use the target article's metric system height measures, rather than translate those to the target sample's imperial system, because we were not sure how such a change might impact results, though expected that if the results in Problem 6c would deviate from the pattern of results in Problem 6a and 6c, this might be one of likely reasons.

Likelihood of sampling outcomes in small vs. large samples (problems 7–9)

KT administered three problems that further tested the representativeness hypothesis concerning sample size. Participants were presented with three problems that involved a sampling process. Each problem has a specified mean of probability, and participants were asked to determine if an outcome that is more/less extreme (between-participants conditions) than a specified critical value is more probable in a larger sample, smaller sample, or the same in both. The following is an example:

> A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower.
>
> For a period of 1 year, each hospital recorded the days on which (more/less) than 60% of the babies born were boys. Which hospital do you think recorded more such days?

Statistically speaking, it is more probable to have an outcome more extreme than the specified critical value (more than 60%) in a smaller sample (the smaller hospital), whereas it is more probable to have an outcome less extreme than the specified critical value (less than 60%) in a larger sample (the larger hospital). This is because the standard error in a larger sample is smaller, and, thus, more likely to average out extremes. Again, as sample size is not representative of the major characteristic of the process from which it originated, KT hypothesised that participants would not have a preference for the correct answer. The target article did not conduct any statistical tests. We evaluated the results based on *p*-values and Bayes Factor.

Posterior probabilities (problems 10 and 11)

KT used two different tasks to measure subjective posterior probabilities: a symmetric binomial task and a non-binomial task.

Binomial task (problem 10). This problem was tested using a 2 x 5 between-participants design in the target article. The problem extended the analysis of sampling distributions to posterior probability judgement by evaluating how subjective posterior probability is affected by the most salient feature: sample proportion. The posterior probability is the probability that a given sample is drawn from one rather than another population. Problem 10 contained 10 conditions that varied the population proportion, sample ratio, and sample difference.

For example, in the following version, the sample ratio is 8:4 (meaning that out of a total of 12 cards drawn from the deck, 8 cards are marked X and 4 cards are marked O), the sample difference is 4 (there are 4 more cards marked X than cards marked O in the sample), and population proportions are 5/6 and 1/6 (5 out of 6 cards are marked X, while 1 out of 6 cards are marked O in the deck A population):

> Consider two very large decks of cards, denoted A and B. In deck A, 5/6 of the cards are marked X, and 1/6 are marked O. In deck B, 1/6 of the cards are marked X, and 5/6 are marked O. One of the decks has been selected by chance, and 12 cards have been drawn at random from it, of which 8 are marked X and 4 are marked O. What do you think the probability is that the 12 cards were drawn from deck A, that is, from the deck in which most of the cards are marked X?

KT hypothesised that participants would rely on the sample proportion of the two objects as this is the most representative feature. More specifically, they hypothesised that in both pairs of population proportions (5/6 and 1/6 vs. 2/3 and 1/3), participants' posterior estimates in the 5:1 sample proportion condition would be larger than in the 4:2, 8:4, and 40:20 conditions, which again would be larger than in the 18:14 conditions. While it is logical that participants perceive a higher likelihood of drawing from Deck A in scenarios with higher sample proportions of X cards, the key point here is that participants do not consider the broader context of the entire population. In other words, people judge likelihood based on what is immediately observed (the sample) rather than considering the broader context (the entire deck).

Non-binomial task (problem 11). Problem 11 examined whether people's tendency to evaluate the posterior probability based on the most salient feature of the sample also applies to non-binomial problems. The problem read as follows:

> The average heights of adult males and females in the US are, respectively, 5 ft 10 in. and 5 ft 4 in. Both distributions are approximately normal with a standard deviation of about 2.5 in. An investigator has selected one population by chance and has drawn from it a random sample. What do you think are the odds that he has selected the male population if:
>
> i. the sample consists of a single person whose height is 5 ft 10 in.?
>
> ii. the sample consists of 6 persons whose average height is 5 ft 8 in.?

Although estimates in (ii) are larger than (i), KT hypothesised that the probability estimates would be larger for (i) than (ii) because the height of the single person matches the population mean. Specifically, because the sample mean is the most salient feature, participants would rely on the similarity of the sample mean to the population mean in their estimation with little regard to sample size. KT noted that although the correct odds are 16 for case (i) and 29 for case (ii), participants judged scenario (i) as more likely, concluding that participants' responses were not only conservative but also violated the correct ordering of likelihoods.

We note that this question included statistical language that we were not sure laypersons in our sample would be able to understand, and in their target sample participants had some kind of statistics background. We decided to follow the target's design and did not make any adjustments because we were not sure how such a change might impact results.

Modifications

We followed the design and procedure of the target article and added a few changes. One major change was made to the number of problems that each subject had to complete. KT explained that participants answered "a small number (typically 2–4) of questions each of which required, at most, 2 min" (p. 432). It would be cognitively demanding for participants to respond to all nine problems. Therefore, to replicate all problems while keeping cognitive demands at a reasonable level, we randomly assigned participants to receive five out of the nine problems. To account for this change, we doubled the sample size (details can be referred to in the supplementary material on the OSF project page).

In addition, we made a change to Problem 6 (Sampling Distributions). In the target article, there were three categories of sampling distributions in this problem, and in each of them, three different sample sizes, adding up to a total of nine conditions with participants randomly assigned to one. In this replication, although we used the same nine conditions, , as noted earlier, participants could be assigned one, two, or all three Problem 6 scenarios within their set of randomly assigned five problems (they were randomly assigned to complete five problems from the total pool, which included these three Problem 6 scenarios). For each scenario, participants were assigned to one of three sample sizes ($N = 10$, $N = 100$, $N = 1,000$).

Finally, we made a minor change to the non-binomial version of the Posterior Probabilities Problem. In the target article, participants were asked to fill in their answers in odds. In this replication, we asked participants to fill in their answers in percentage, which should be easier to understand. This aligns with the first version of the problem (the symmetrical binomial version), which also requested participants to fill in their answers in percentages.

Extensions

Decision Styles Scale. As an extension to the current replication, we measured participants' decision styles (i.e., preference for intuitive and analytical reasoning) using the decision styles scale (DSS; Hamilton et al., 2016). The scale contains two dimensions; one reflecting an analytical style (five items) and another reflecting an intuitive style (five items). Participants rated their agreement with statements on a 5-point Likert-type scale (1 = *Strongly disagree*, 5 = *Strongly agree*). Example items include "I prefer to gather all the necessary information before committing to a decision" (rational dimension) and "When making decisions, I rely mainly on my gut feelings" (intuitive dimension). The scales demonstrated strong reliabilities ($\alpha = .88$ for both subscales). The full scale is available in the online Supplementary Material Table S3.

Extent of using the representativeness heuristic. We calculated the number of answers using the representativeness heuristic divided by the number of problems attempted. The dependent variable varies from 0 (no use of the representativeness heuristic) to 1 (full use of the representativeness heuristic).

We summarised the options that indicate using the representativeness heuristic in Table 6. In the pre-registration, we initially specified that five of the nine problems scored the use of the representative heuristic and thus planned to

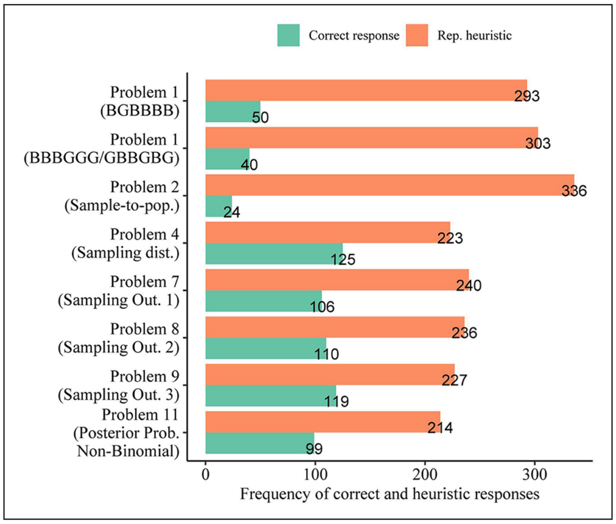


Figure 1. Frequency of heuristic responses.

include only those. These were: Problem 2 (sample-to-population, high-school programme), Problem 4 (reflection of randomness), and Problem 7 (likelihood of sampling outcomes, babies scenario). We later realised that other problems scored the representativeness heuristic too and thus decided to include them in calculating the extent of the representativeness heuristic (see Figure 1). As noted earlier, given that participants randomly received five out of nine problems, some of them did not complete some of the problems that involved the representativeness heuristic.

Evaluation of replication closeness

Based on the criteria set by LeBel et al. (2018), we classified the replication as a close replication (see Table 4). We attempted to replicate nine problems from the target article. We also aimed to compare the replication effects with the original effects in the target article using the criteria set by LeBel et al. (2019): (1) whether a signal was detected (i.e., whether the confidence interval includes 0); (2) the consistency of the replication effect size estimate with that observed in the original study (i.e., whether the replication's CI includes the original effect size point estimate); and (3) the magnitude of the replication's effect size estimate in the same direction compared to original effect size. For the two posterior probabilities problems (Problems 10 and 11), effect sizes could not be calculated due to insufficient information provided in the target article.

Table 4. Classification of the replication, based on LeBel et al. (2018).

Design facet	Replication	Details of deviation
IV operationalization	Same	/
DV operationalization	Same	/
IV stimuli	Same	/
DV stimuli	Same	/
Procedural details	Similar	The details can be referred to in the "design and procedure" section in this manuscript.
Physical settings	Different	The target article conducted the experiment in a classroom setting with pen and paper, whereas the replication was conducted online.
Contextual variables	Different	Same as above.
Replication classification	Close replication	

Deviations from pre-registration

We note that we deviated from the pre-registration plan in several ways.

First, we reported the results with (pre-registered) exclusions in the main manuscript and reported pre-exclusion results in the online Supplementary Material. We did this to ensure that we excluded participants who might not have been able to stay concentrated on the experiment, given its cognitively demanding nature.

Second, in the pre-registration, we specified that five of the nine problems scored the use of the representative heuristic, and thus only planned to include those. However, we later realised that other problems also provided a way to score the representativeness heuristic, and therefore extended the analysis to include those as well. This is further detailed in the section "Extent of Using the Representativeness Heuristic."

Problems 7–9 compared participants' responses against random chance. We conducted two additional tests. First, we added tests using a different expected proportion, and we conducted the equivalence test using a different method to set the smallest effect size of interest.

Regarding decision styles, we pre-registered that we would treat decision style as a unidimensional measure (e.g., ranging from intuitive to rational). However, this was not accurate as the scale composed of two distinct subscales: intuitive style and rational style. We also pre-registered a linear regression model to test the association between decision style and use of the representativeness heuristic. We kept this analysis but also added a mixed effects model treating "problem" as a within-subject factor/repeated measure, and also explored the interactive influence of both decision styles.

Results

Replication findings overview

We summarised the replication results in Table 5. We summarised the descriptive statistics from the problems that

Table 5. Replication: summary of results.

Problem	Replication summary
1.1 Sample-population similarity (birth sequence)	Successful
1.2 Sample-population similarity (high school programme)	Successful
4. Reflection of randomness	Successful
6. Sampling distributions	Successful
7. Likelihood of sampling outcomes (babies)	Successful
8. Likelihood of sampling outcomes (investigator)	Successful
9. Likelihood of sampling outcomes (disease)	Successful.
10. Posterior probabilities (binomial)	Unsuccessful
11. Posterior probabilities (non-binomial)	Successful

scored the representativeness heuristic in Table 6 and plotted the frequency of heuristic responses in these problems in Figure 1. We provided details regarding each of the problems in the following sections.

We summarised the findings for Problems 1, 2, and 4 in Table 7.

Problem 1 (sample-to-population similarity, birth sequence)

Consistent with the target article, most participants selected the heuristic response.

In Problem 1 (birth sequence), one-proportion z-tests indicated that most participants (293 out of 343) estimated the birth sequence BGBBBB to be less probable than the birth sequence GBGBBG, $\chi^2 = 170.74$, $p < .001$, $h = 0.79$, 95% CI [0.68, 0.89]. Most participants (303 out of 343) estimated the birth sequence BBBGGG to be less probable than the birth sequence GBBGBG,$\chi^2 = 200.13$, $p < .001$, h = 0.87, 95% CI [0.77, 0.98].

Problem 2 (sample-to-population similarity, high-school programme)

In Problem 2 (high school programme), we conducted a one-proportion z-test and found that most participants (336 out of 343 participants) estimated that the class belonged to programme A rather than programme B, $\chi^2 = 268.67$, $p < .001$, $h = 1.05$, 95% CI [0.95, 1.15]. We concluded that our findings are consistent with the target article's.

Problem 4 (reflection of randomness)

We conducted a one-proportion z-test and found that most participants (223 out of 343 participants) estimated distribution I (the non-uniform distribution) to be more probable than distribution II (the uniform distribution), $\chi^2 = 27.04$, $p < .001$, h = 0.28, 95% CI [0.77, 0.98].

Problem 6 (sampling distributions)

We summarised the comparison of the statistical details between the replication and the original findings in Table 8. We plotted participants' mean probability estimates in the three scenarios in Figure 2. All three distributions were consistent with the target article's findings.

The target article did not conduct any statistical tests for this problem. We conducted a Kolmogorov-Smirnov test on each comparison of sample size ($N = 10$ vs. $N = 100$, $N = 100$ vs. $N = 1{,}000$, and $N = 10$ vs. $N = 1{,}000$) in each category of the sampling distribution (distribution of gender, blood type, and height). We did not find evidence for differences in mean probability estimates between sample size conditions in any of the categories. These results are consistent with the target article. We could not quantify the null as we found no Bayesian approach for Kolmogorov–Smirnov tests.

Problems 7–9 (likelihood of sampling outcomes)

We summarised the comparison of the statistical details between the target article and the replication for Problems 7–9 in Table 9 (the three likelihood of sampling outcomes problems).

We ran a series of one-proportion z-tests (one-tailed) for each scenario (babies, investigator, disease) and for each condition ("more extreme" vs. "less extreme") that compared participants' responses against the expected proportion by chance. Following the pre-registration, we set the expected proportion at 33.33% (1/3).

As these three problems tested a null hypothesis, we also used equivalence testing and Bayesian analysis to quantify evidence in favour of the null hypothesis over the alternative hypothesis (Lee & Wagenmakers, 2005).

Most results were in line with the target article's findings, apart from the "More extreme" condition in Problem 8 and both conditions in Problem 9, where the Bayes Factors indicated very strong to extreme evidence for the alternative hypothesis over the null hypothesis (Lee & Wagenmakers, 2014).

Next, we set the expected proportion at 50% (not pre-registered). This is a less conservative test but is arguably more in line with the target article. KT tested whether there was a "significant preference for the correct answer," which we on closer reading interpreted as whether the proportion of correct answers was higher than 50%. Although KT also reported that "About the same" was the modal answer, this is not a statistical test. Moreover, as Teigen (2022) points out, "To test the difference between participants choosing (a) and "equally likely" makes no sense as

Table 6. Replication: descriptive statistics for problems that scored the representativeness heuristic.

Problems		Option	Count	*N*
1: Sample-to-population similarity (birth sequence*)*	Birth sequence BGBBBB	Less than 72	293	343
		Equal to or more than 72*	50	
	Birth sequence BBBGGG vs GBBGBG	BBBGGG equally or more probable*	40	343
		GBBGBG more probable	303	
2: Sample-to-population similarity (high-school programme)		Programme A	336	360
		Programme B*	24	
4: Sampling distributions		Distribution I (non-uniform)	223	348
		Distribution II (uniform)*	125	
7–9: Likelihood of sampling outcomes				
7: *Babies*	More extreme	About the same	71	173
		The smaller hospital*	52	
		The larger hospital	50	
	Less Extreme	About the same	71	173
		The smaller hospital	48	
		The larger hospital*	54	
8: *Investigator*	More extreme	About the same	65	173
		The line investigator*	71	
		The page investigator	37	
	Less Extreme	About the same	67	173
		The line investigator	67	
		The page investigator*	39	
9: *Disease*	More extreme	About the same	52	173
		The team checking 1*	33	
		The team checking 3	88	
	Less extreme	About the same	55	173
		The team checking 1	32	
		The team checking 3*	86	

Note. Correct answers (no use of representativeness heuristic) are starred.

Table 7. Problems 1 and 2 (sample-to-population similarity) and problem 4 (reflection of randomness): comparison of the findings in target article versus replication.

		Target article		Replication		
Problem		*P*	Cohen's *h* [95% CI]	*p*	Cohen's *h* [95% CI]	Interpretation
1. Sample-Population Similarity (birth sequences)	BGBBBB vs GBGBBG	$<.001$	0.68 [0.48, 0.89]	$<.001$	0.79 [0.68, 0.89]	Signal–consistent
	BBBGGG vs GBBGBG	$<.001$	/	$<.001$	0.87 [0.77, 0.98]	Signal NA (effect size of the target article is not available)
2. Sample-Population Similarity (high school programmes)		$<.001$	0.53 [0.32, 0.74]	$<.001$	1.05 [0.95, 1.15]	Signal–inconsistent, larger
4. Reflection of Randomness		.007	0.39 [0.12, 0.67]	$<.001$	0.29 [0.18, 0.39]	Signal–consistent

Note. Sign tests were conducted in the target article and one-proportion z-tests in the replication.

no meaningful null hypothesis can be formed." (p. 193). With this 50% as the expected proportion, the results are consistent with the target article. That is, the number of participants choosing the correct answer did not exceed 50% in any of the problems.

To further quantify the null in Problems 7–9 (Likelihood of Sampling Outcomes), we examined whether the confidence intervals of each effect in the replication contained the smallest effect size of interest (SESOI). As per the preregistration, we specified the SESOI by halving the

Table 8. Problem 6 (sampling distributions): comparison of findings in target article versus replication.

		Target article	Replication	
Categories of sampling distributions	Comparisons of sampling distributions	*p*	*D*	*p*
Distribution of genders	*N* = 10 vs *N* = 100	1.00	0.18	.997
	N = 100 vs *N* = 1,000	.993	0.36	.479
	N = 10 vs *N* = 1,000	.993	0.27	.833
Distribution of blood type	*N* = 10 vs *N* = 100	.993	0.18	.997
	N = 100 vs *N* = 1,000	.993	0.46	.211
	N = 10 vs *N* = 1,000	.993	0.36	.479
Distribution of height	*N* = 10 vs *N* = 100	1.00	0.14	1.00
	N = 100 vs *N* = 1,000	.938	0.14	1.00
	N = 10 vs *N* = 1,000	.938	0.14	1.00

Note. We conducted Kolmogorov-Smirnov tests on the given data of the target article. Kolmogorov-Smirnov tests were also conducted in the replication. *D* is the effect size for Kolmogorov-Smirnov tests.

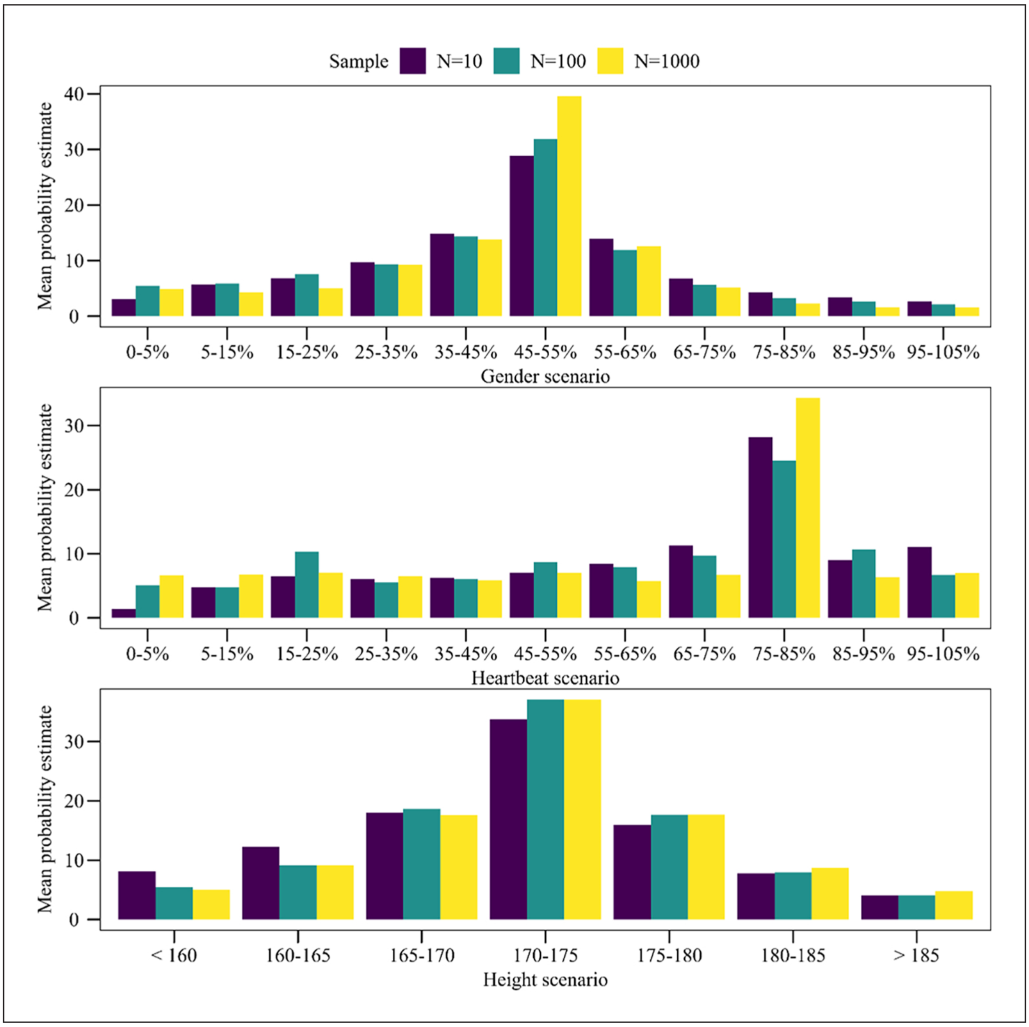


Figure 2. Problem 6 (sampling distributions): mean probability estimates of sampling distributions.

smallest effect size in the previous problems in the target article (Problems 1, 2, and 4). Problem 4 in the target article had the smallest effect (Cohen's h = 0.39, 95% CI = 0.12, 0.67). Halving this effect size resulted in a Cohen's h of about 0.20 (95% CI = 0.06, 0.33). We interpreted effects below the lower confidence interval of the SESOI as

Table 9. Problems 7–9 (likelihood of sampling outcomes): statistical tests.

		Target article		Replication			
Problem	Condition	*p*	Cohen's *h* [95% CI]	*p*	Cohen's *h* [95% CI]	BF_{10} (BF_{01})	Replication summary
7 ("Babies")	More extreme	.064	-0.30 [-0.58, -0.03]	.798	-0.07 [-0.22, 0.08]	0.43 (2.30)	No signal– inconsistent, smaller
	Less extreme	.082	-0.30 [-0.60, -0.01]	.695	-0.04 [-0.19, 0.10]	0.68 (1.47)	No signal– inconsistent, smaller
8 ("Investigator")	More extreme	.207	0.20 [-0.08, 0.48]	.019	0.16 [0.01, 0.31]	64.93 (0.02)	Signal– consistent
	Less extreme	.092	-0.28 [-0.20, 0.37]	.998	-0.24 [-0.39, -0.09]	0.55 (1.82)	No signal– consistent
9 ("Disease")	More extreme	.646	0.09 [-0.20, 0.37]	1	-0.33 [-0.48, -0.18]	9,102 (0.000)	No signal– inconsistent, opposite
	Less extreme	.646	-0.09 [-0.37, 0.19]	<.001	0.33 [0.18, 0.48]	1,819 (0.001)	Signal– inconsistent, opposite

Note. One-proportion z-test (one-tailed in the replication), $N = 346$, $N_{\text{More extreme}} = 173$, $N_{\text{Less extreme}} = 173$. BF = Bayes factor, quantifying evidence for the alternative (BF_{10}) and the null (BF_{01}). Two-tailed *p*-values for the target article and one-tailed *p*-values in the replication. "Smaller" means that the effect is closer to zero.

Table 10. Problems 10 and 11 (posterior probabilities problems): subjective probability estimates.

	Target article		Replication		
	n	*M*	*n*	*M*	*SD*
Binomial problem: (format: Initial proportion in decks, sample proportion)					
2:3, 18:14	56	58	35	70.43	12.93
2:3, 4:2	56	68	36	68.39	10.93
2:3, 8:4	56	70	39	68.92	11.87
2:3, 40:20	56	70	35	77.34	16.91
2:3, 5:1	56	85	31	77.45	13.26
5:6, 18:14	56	60	32	71.12	15.54
5:6, 4:2	56	70	35	76.60	13.78
5:6, 8:4	56	70	34	68.59	14.14
5:6, 40:20	56	70	34	76.26	15.37
5:6, 5:1	56	83	35	85.37	15.04
Non-binomial problem					
type (i)	115	88.89	347	65.90	26.06
type (ii)	115	71.43	347	56.59	26.03

Note. Subjective Probability Estimates are expressed as percentages. *n* for the binomial problem in the original is the average number of participants in that condition. KT reported that the number of participants for each of the 10 conditions in this problem ranged from 37 to 79, with an average of 56.

practically equivalent to zero. Only the effect in the less extreme condition in Problem 7 was lower than the lower confidence interval of the SESOI, suggesting that for the remaining effects, we cannot conclude the absence of an effect.

Next, we conducted an exploratory equivalence test with a less conservative and more common approach. Specifically, we used Simonsohn's (2015) small-telescope approach and defined the SESOI as the effect size the target article had 33% power to detect. This was not pre-registered. With this approach, the SESOI was $h = 0.16$. All effects had confidence intervals that included 0.16, suggesting that we cannot conclude the absence of an effect.

Problems 10 and 11 (posterior probabilities)

We summarised the descriptives for Problems 10 and 11 in Table 10. We summarised the comparison of the statistical details between the target article's and replication's findings for Problems 10 and 11 (the two posterior probabilities problems) in Table 11.

We conducted a one-way ANOVA and Tukey post hoc tests on target comparisons. Recall that KT hypothesised that people would rely on the sample proportion as this is the most representative feature. Specifically, they hypothesised that for both pairs of population proportions (5/6 and 1/6 vs. 2/3 and 1/3), participants' posterior estimates in the 5:1 sample proportion condition would be larger than in the 4:2, 8:4, and 40:20 conditions, which again would be larger than in the 18:14 conditions. This is non-normative: for example, the 40:20 sample provides much stronger evidence than the 5:1 sample. For the conditions with the initial proportion of 2:1 in the deck, we found that the posterior probabilities stated by the participants in conditions 5:1 had no difference from the ones in 4:2, 8:4, and 40:20.

Next, the posterior probabilities stated by the participants in conditions 18:14 also had no difference with the ones in 4:2, 8:4, and 40:20. For the conditions with the initial proportion of 5:1 in the deck, we found that the posterior probabilities stated by the participants in conditions 5:1 had no difference with the ones in 4:2 and 40:20, but were larger than the ones in condition 8:4.

Table 11. Problems 10 and 11 (posterior probabilities): comparison of target article and replication.

		Target article			Replication		
		t	*p*	Cohen's *d* [95% CI]	*t*	*p*	Cohen's *d* [95% CI]
Binomial problem							
Initial proportion in the decks	Comparison of different sample proportion						
2:1	5:1 vs 4:2	/	<.01	/	2.6340	.009	0.64 [0.15, 1.13]
	5:1 vs 8:4	/	<.01	/	2.5240	.012	0.60 [0.12, 1.08]
	5:1 vs 40:20	/	<.01	/	0.0314	.975	0.01 [-0.47, 0.49]
	18:14 vs 4:2	/	<.01	/	0.612	.541	0.15 [-0.32, 0.61]
	18:14 vs 8:4	/	<.01	/	0.460	.645	0.11 [-0.35, 0.56]
	18:14 vs 40:20	/	<.01	/	-2.060	.040	-0.49 [-0.97, -0.02]
5:1	5:1 vs 4:2	/	<.01	/	2.613	.009	0.62 [0.14, 1.10]
	5:1 vs 8:4	/	<.01	/	4.9634	<.001	1.20 [0.68, 1.70]
	5:1 vs 40:20	/	<.01	/	2.6932	.007	0.65 [0.16, 1.13]
	18:14 vs 4:2	/	<.01	/	-1.594	.112	-0.39 [-0.87, 0.09]
	18:14 vs 8:4	/	<.01	/	0.733	.464	0.18 [-0.30, 0.66]
	18:14 vs 40:20	/	<.01	/	-1.4861	.138	-0.37 [-0.85, 0.12]
Non-binomial problem		/	<.01	/	6.19	<.001	0.33 [0.22 0.44]

Note. For the binomial problem, median tests were conducted in the target article, whereas one-way ANOVA with pairwise comparisons was conducted in the replication (*N*=346). For the non-binomial problem, a median test was conducted in the target article, whereas a paired sample *t*-test was conducted in the replication (*N*=347).

Next, the posterior probabilities stated by the participants in condition 18:14 also were not different from the ones in 4:2, 8:4, and 40:20 conditions.

The target article found that estimated posterior probabilities in conditions 5:1 were larger than those in 4:2, 8:4, and 40:20 for both sets of initial probabilities. Also, estimated posterior probabilities in conditions 18:14 were smaller than those in 4:2, 8:4, and 40:20 for both sets of initial probabilities. However, in the replication, only the estimated posterior probabilities in condition 5:1 were larger than those in 8:4 for the initial probability of 5:1. We did not find evidence for differences in the remaining comparisons. Nevertheless, similar to KT, we found that participants were insensitive to population proportions.

For Problem 12 (posterior probabilities, non-binomial), we conducted a paired-sample *t*-test and found that participants attached greater probability to selecting the male population if the sample consisted of a single person whose height was 5 ft 10 in. (case *[i]*) than if the sample consisted of 6 persons whose average height was 5 ft and 8 in. (case *[ii]*), t=6.19, p<.001, d=0.33, 95% CI [0.24, 0.47], which is opposite to the normatively correct answer. Our replication results were very similar to those of the target article.

Extension: decision style

As an extension to the replication, we examined if the decision styles correlated with the extent of using the representativeness heuristic. We calculated reliance on the representativeness heuristic by taking the ratio of scores in Problems 1.1, 1.2, 2, 4, 7, 8, and 9 to the number of heuristic-scoring problems they completed, ranging from 0 to 1 (M=0.75, SD=0.23, Med=0.75). In our pre-registration, we omitted Problems 1.1, Problem 1.2, and Problem 11 from the calculation because we did not initially recognise that these problems also scored the representativeness heuristic.

We did not find support for the hypothesis that reliance on the representativeness heuristic correlates with intuitive (r=0.03, p=.422, 95% CI=-0.05, 0.11) or rational decision style (r=0.03, p=.524, 95% CI=-0.05, 0.10). Neither did it correlate with age (r=.03, p=.411, 95% CI=-0.05, 0.11), gender (r=-.00, p=1, 95% CI=-0.08, 0.08), or education (r=-.02, p=.579, 95% CI=-0.10, 0.06).

We next examined these associations in a binomial mixed effects model that included "problem" and "subject" as random factors, using the *lme4* package in *R* (Bates et al., 2014). We restructured the data to long format and treated problem as a repeated measure (not pre-registered). We did not find an association between the intuitive (B=0.04, p=.460, 95% CI=-0.07, 0.16) nor the rational style (B=0.15, p=.113, 95% CI=-0.04, 0.33) with the representativeness heuristic.

As an additional exploratory analysis, we examined whether the two styles interactively predicted reliance on the representativeness heuristic. Dual-process theorists suggested that the two styles are conceptually independent and operate interactively (Kahneman, 2002; Norris &

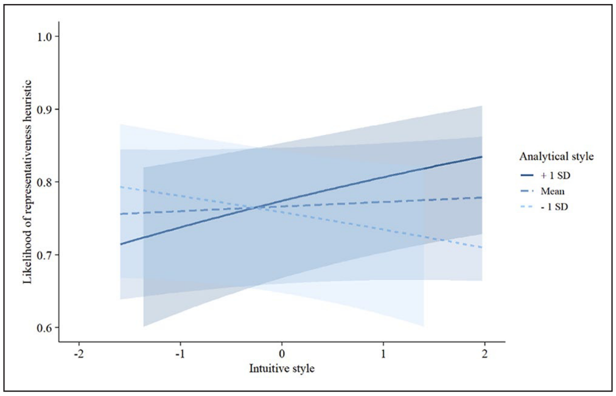


Figure 3. Interaction between intuitive and rational styles in predicting representativeness heuristic.
Note. Predictors are mean-centred.

Epstein, 2011; Stanovich & West, 2000). Thus, individuals can be grouped into four different categories: high on both styles, low on both styles, high on rationality and low on intuition, and low on rationality and high on intuition (Epstein, 1998; Bakken et al., in press; Hodgkinson & Clarke, 2007; Hodgkinson et al., 2009; Shiloh et al., 2002).

We found a cross-over interaction ($B = 0.26$, $p = .009$, 95% CI = 0.07, 0.46), which we plotted in Figure 3. The interaction plot suggests that those who were high on both dimensions were more prone to using the representativeness heuristic, which is consistent with previous findings (e.g., Shiloh et al., 2002). The results were similar in the pre-exclusion analysis (see the online Supplementary Material). We will return to these findings in the Discussion.

Associations and comparisons between problems

One notable strength of the current replication study is that participants completed multiple problems, in contrast to the target article where each problem was presented to a different sample. This setup enabled us to assess the consistency of heuristic responses across problems.

First, we examined the correlations between responses in all of the heuristic-scoring problems (Table 12). We only found evidence for a positive correlation between Problems 7 and 9 and a negative correlation between Problems 4 and 8. These results suggest very poor consistency in participants' responses to the problems.

Next, we explored pairwise comparisons between all problems. We used the *lme4* package (Bates et al., 2014) and ran a logistic mixed effects model with heuristic response (0 = non-heuristic response, 1 = heuristic response) as the dependent variable, problem as the independent variable, and subject as a random factor. The pairwise comparisons using Tukey's test are plotted in Figure 4. Results are given on the log odds ratio scale.

Figure 4 indicates that Problem 1 (sample-to-population similarity, birth sequence) differed from almost all of the other problems. Problem 2 (sample-to-population similarity, high-school programme) differed slightly from Problem 1, but more from Problems 4–11. We found no support for pairwise comparisons differences among Problems 4–11. A visual inspection of these pairwise comparisons suggests two clusters of problems.

Discussion

We conducted a pre-registered replication of Kahneman and Tversky's (1972) classic article on the representativeness heuristic. As the target article included a relatively large set of problems with a variety of different contexts, it is perhaps unrealistic to expect that the pattern of results observed by Kahneman and Tversky in the 1970s will be replicated 50 years later. Yet, for all but one of the problems, our replication results are remarkably similar to the target article's.

Problems 1 and 2 (sample-population similarity) were successfully replicated. In Problem 1 (birth sequence), participants seemed to be sensitive to the similarity of a sequence of births to (a) the proportion of cases in the population and (b) the order of events, with "streaks" of three boys and three girls seen as non-random and thus less probable. Similar to (a), in Problem 2 (high school programme), people seemed sensitive to the similarity in proportions or majority/minority relation between a sample and a population.

Problem 4 (reflection of randomness) was also successfully replicated, and similar to (b), indicated that people have ideas about how random sequences or distributions "should" look, with a uniform distribution thought to be too orderly to be really random, and thus less probable than a nonuniform distribution.

Problem 6 (sampling distributions) addresses a slightly different concern, namely (c) insensitivity to sample size. The target article's results were successfully replicated, with people seemingly relying too much on the salient feature of the sample proportion or mean, and essentially ignoring sample size, leading them to suggest much too wide distributions for large samples.

The issue of (in)sensitivity to sample size was also addressed in Problems 7–9 (the three likelihood of sampling outcomes problems). Here, it is possible to argue for different interpretations of whether the original findings were replicated. Participants were tasked to judge whether a sampling outcome more or less extreme than a specified value is more likely to occur in a small or a large sample, or about equally likely. KT argued that people would be insensitive to sample size, and reported that the modal answer was equally likely in "almost all comparisons" (5 out of 6). In the replication, the modal answer was equally likely in only 3 out of 6 comparisons. Nonetheless, as in

Table 12. Heuristic response problems: correlations.

P#	1.1	1.2	2	4	7	8	9
1.2	-.02 [-0.13, 0.08] (343)						
2	-.11 [-0.25, 0.04] (180)	.06 [-0.09, 0.20] (180)					
4	.09 [-0.06, 0.24] (167)	-.07 [-0.22, 0.08] (167)	-.10 [-0.24, 0.04] (185)				
7	-.01 [-0.16, 0.15] (169)	-.02 [-0.17, 0.13] (169)	-.02 [-0.16, 0.12] (204)	-.01 [-0.17, 0.14] (163)			
8	-.13 [-0.28, 0.02] (169)	-.03 [-0.18, 0.12] (169)	.02 [-0.12, 0.15] (204)	-.16* [-0.31, -0.01] (163)	.08 [-0.02, 0.19] (346)		
9	.05 [-0.10, 0.20] (169)	-.01 [-0.16, 0.14] (169)	.08 [-0.06, 0.21] (204)	.05 [-0.11, 0.20] (163)	.23** [0.13, 0.33] (346)	-.10 [-0.21, 0.00] (346)	
11	-.07 [-0.22, 0.09] (160)	.04 [-0.12, 0.19] (160)	-.07 [-0.21, 0.09] (170)	-.11 [-0.26, 0.05] (161)	.04 [-0.12, 0.20] (150)	.05 [-0.11, 0.21] (150)	.01 [-0.15, 0.17] (150)

*$p < .05$. **$p < .01$.

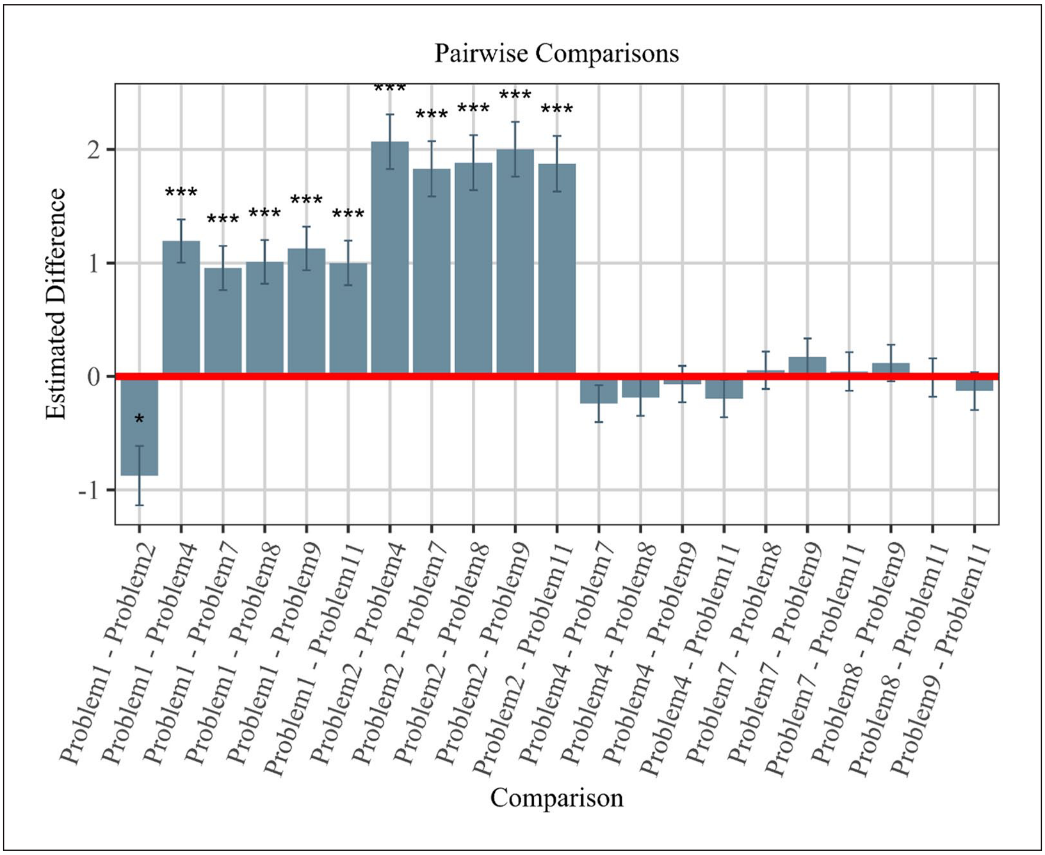


Figure 4. Heuristic response problems: pairwise comparisons.
Note. Problem 1 in the figure refers only to the first sub-question (Problem 1.1). The second sub-question (Problem 1.2), which is very similar and included mainly as a robustness check, is not included in this analysis.

the target article, we did not find a systematic preference for the correct answer. Even in the best performing condition (people who judged the likelihood of the less extreme outcome in the disease scenario), only 86 out of 173 participants (49.7%) chose the correct outcome, with a (very slight) majority choosing incorrect options. Thus, the replication results support the idea that people are insensitive to sample size, but if one argues that the representativeness heuristic would predict that "equally likely" would always be the modal answer, results are more mixed.

Problems 10 and 11 (posterior probabilities, binomial and non-binomial) investigated yet another kind of probability judgement. Problem 10 (binomial posterior probability) concerned the probability of a sample being taken from one of two populations (two different decks of cards with different proportions of cards marked with X's and O's), given different samples with different ratios of X's and O's, and differences of X's and O's. KT found that people relied strongly on the sample ratio, with little concern for sample differences. While the replication results do not match the original results when it comes to the exact ordering of probabilities and do not show differences between judged probabilities in the same fashion as KT, the results are consistent in the sense that the subjective probabilities do not follow normative rules. The 5:1 observed ratio is still given the highest subjective probability of coming from the target population, even though the 40:20 sample provides much stronger evidence. Thus, although the original results are not replicated, again the replication results show non-normative probability judgements and indicate that the sample ratio is given more weight than it possibly should as compared to the sample difference.

Problem 11 (non-binomial posterior probability) was successfully replicated: Participants were entirely insensitive to sample size and seemed to base their judgements on the similarity of the sample mean to the population mean, leading to an incorrect ordering of likelihoods.

Overall, the results indicate that the representativeness heuristic is alive and well, at least in the sense that we found similar results as the target article in most of the problems. Even in those problems where it can be debated whether the target article's findings were replicated, our results show that subjective probability judgements using these problems do not follow normative rules, but are based on subjective impressions and arguably consistent with representativeness playing a role. Notwithstanding the long-lasting controversy about the vagueness of representativeness as a theoretical concept (Gigerenzer, 1996; Teigen, 2022), these results indicate that the target article's findings seem to hold up well and that the debate can proceed with discussions of how to interpret the findings rather than questioning their robustness.

Furthermore, our study contributes to the literature by examining the internal consistency and convergence of responses across multiple problems that tap into the representativeness heuristic. While previous studies have typically focused on between-subjects designs, our inclusion of multiple problems completed by the same participants allows us to assess the coherence of different conceptualizations of the representativeness heuristic. Very few replications address such a wide range of tests of the same underlying concept using a within-subjects approach.

The exploratory analyses showed that reliance on the representativeness heuristic varied considerably across problems, suggesting that responding in a representativeness-based way for one problem does not mean that individuals will base their judgement on representativeness for a different problem. Pairwise comparisons among all of the problems also indicated support for half of the comparisons, suggesting that several of the problems differed with respect to predicting reliance on the representativeness heuristic. With this in mind, one could question whether the target article has collected a wide range of problems that may not tap into the same mechanism, or at least that different people are differently prone to base their judgements on representativeness in different situations. Our findings align with the idea that even the same bias or heuristic might derive from different processes and might depend differently on various individual differences (Ceschi et al., 2019). These findings also align with the results from recent efforts testing the reliability of judgement and decision-making tasks (e.g., anchoring; Röseler et al., 2022), and address recent calls to test the reliability of cognitive behavioural tasks (e.g., Parsons et al., 2019).

Finally, we extended the replication by examining the relationship between decision styles and the extent of using the representativeness heuristic. According to Kahneman and Tversky, heuristics are driven by automatic intuitive responses, which can be overridden through deliberate processing (Kahneman & Frederick, 2002). We thus hypothesised that the extent of using the representativeness heuristic would be positively correlated with an intuitive decision style and negatively correlated with a rational decision style. We did not find evidence for this hypothesis.

Nevertheless, we found some evidence for a possible interaction between the two styles, which is consistent with previous research. Shiloh et al. (2002) observed the same cross-over interaction between the two styles in predicting susceptibility to framing effects. Shiloh and colleagues speculated that those who scored high or low on both styles are more sensitive to environmental cues and therefore also more sensitive to framing:

> In order to be resistant to framing effects, individuals should have a clearly dominant thinking style, either rational or intuitive. Both have strong internal guides, either logical or experiential, upon which they rely in processing information in risky situations. However, people with non-differentiated

> thinking styles [. . .] tend to rely more on cues within the situation, rendering them more susceptible to biases like framing effects (p. 425).

Finally, there are limitations in the reliability of the extent of using the representativeness heuristic. We calculated this by taking the ratio of participants' scores in the problems that scored the representativeness heuristic to the number of problems that they completed. However, participants were randomly assigned to complete five out of nine problems (due to the high cognitive demand of the survey) and one of the problems did not score reliance on the representativeness heuristic. These variations in problem assignment across participants of the study create certain unreliability in the heuristic variable. Future research may want to focus on fewer problems or ensure that each participant completes the same number of problems.

Conclusion

Our replication of Kahneman and Tversky's (1972) seminal article on the representativeness heuristic underscores its enduring influence and the robustness of their findings. Notably, however, heuristic responses varied across problems. Further research is needed to elucidate the underlying mechanisms influencing individuals' reliance on this heuristic.

Authorship declaration

Kai Hin Wan conducted the replication as part of his thesis in psychology under the supervision of Gilad Feldman.

Lewend Mayiwar revised, verified, reproduced all analyses and added new analyses, and wrote the initial journal submission manuscript. Erik Løhre guided Lewend and gave feedback on drafts.

Gilad Feldman guided the replication effort, supervised each step in the project, ran data collection, conducted the pre-registration, and edited the manuscript for submission.

Author contributions

Lewend Mayiwar: Formal analysis; data analysis peer review/verification; validation; visualisation; writing—original draft; writing—review and editing.

Wan Kai Hin: Conceptualisation; pre-registration; data curation; formal analysis; investigation; methodology; software; validation; visualisation; writing—original draft; writing—review and editing.

Erik Løhre: Supervision; writing—review and editing.

Gilad Feldman: Conceptualisation; pre-registration; data curation; funding acquisition; investigation; pre-registration peer review/verification; project administration; resources; software; supervision; writing—review and editing.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project has been supported by the Teaching Development Grant from the University of Hong Kong awarded to Gilad Feldman.

ORCID iD

Gilad Feldman https://orcid.org/0000-0003-2812-6599

Data accessibility statement

The data and materials from the present experiment are publicly available at the Open Science Framework website: https://osf.io/nhqc4/

Supplementary material

The Supplementary Material is available at: qjep.sagepub.com

References

Agnoli, F. (1991). Development of judgmental heuristics and logical reasoning: Training counteracts the representativeness heuristic. *Cognitive Development*, *6*(2), 195–217. https://doi.org/10.1016/0885-2014(91)90036-D

Alaybek, B., Wang, Y., Dalal, R. S., Dubrow, S., & Boemerman, L. S. (2022). The relations of reflective and intuitive thinking styles with task performance: A meta-analysis. *Personnel Psychology*, *75*(2), 295–319. https://doi.org/10.1111/peps.12443

Au, N., & Feldman, G. (2020). *Revisiting "goals as reference points": Replication and extensions of Heath et al. (1999)*. https://doi.org/10.17605/OSF.IO/WMQTB

Bakken, B. T., Hansson, M., & Hærem, T. (2024). Challenging the doctrine of "non-discerning" decision-making: Investigating the interaction effects of cognitive styles. *Journal of Occupational and Organizational Psychology*, *97*, 209–232. https://doi.org/10.1111/joop.12467

Bar-Hillel, M. (1984). Representativeness and fallacies of probability judgment. *Acta Psychologica*, *55*(2), 91–107. https://doi.org/10.1016/0001-6918(84)90062-3

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *ArXiv:1406.5823*. http://arxiv.org/abs/1406.5823

Brannon, L. A., & Carson, K. L. (2003). The representativeness heuristic: Influence on nurses' decision making. *Applied Nursing Research*, *16*(3), 201–204. https://doi.org/10.1016/S0897-1897(03)00043-0

Busenitz, L. W., & Barney, J. B. (1997). Differences between entrepreneurs and managers in large organizations: Biases and heuristics in strategic decision-making. *Journal of Business Venturing*, *12*(1), 9–30. https://doi.org/10.1016/S0883-9026(96)00003-1

Ceschi, A., Costantini, A., Sartori, R., Weller, J., & Di Fabio, A. (2019). Dimensions of decision-making: An evidence-based classification of heuristics and biases. *Personality* *and Individual Differences*, 146, 188–200. https://doi-org.ezproxy.library.bi.no/10.1016/j.paid.2018.07.033

Chan, H. C., & Feldman, G. (2024). *Representativeness heuristic in intuitive predictions: Replication Registered Report of problems reviewed in Kahneman and Tversky (1973)*. https://doi.org/10.17605/OSF.IO/8ZHCJ

Chandrashekar, S., Cheng, Y. H., Fong, C. L., Leung, Y. C., Wong, Y. T., Cheng, B. L., & Feldman, G. (2021). Frequency estimation and semantic ambiguity do not eliminate conjunction bias, when it occurs: Replication and extension of Mellers, Hertwig, and Kahneman (2001). *Meta-Psychology*, *5*. https://doi.org/10.15626/MP.2020.2474

Chatterjee, S., Heath, T. B., Milberg, S. J., & France, K. R. (2000). The differential processing of price in gains and losses: The effects of frame and need for cognition. *Journal of Behavioral Decision Making*, *13*(1), 61–75. https://doi.org/10.1002/(SICI)1099-0771(200001/03)13:1%3C61::AID-BDM343%3E3.0.CO;2-J

Cox, C., & Mouw, J. T. (1992). Disruption of the representativeness heuristic: Can we be perturbed into using correct probabilistic reasoning? *Educational Studies in Mathematics*, *23*(2), 163–178. https://doi.org/10.1007/BF00588054

Epstein, S. (1998). Cognitive-experiential self-theory. In D. F. Barone, M. Hersen & V. B. Van Hasselt (Eds.), *Advanced personality. The Plenum series in social/clinical psychology* (pp. 211–238). Springer. https://doi.org/10.1007/978-1-4419-8580-4_9

Feldman, G. (2023). *Registered Report Stage 1 manuscript template*. https://doi.org/10.17605/OSF.IO/YQXTP

Fuster, A., Laibson, D., & Mendel, B. (2010). Natural expectations and macroeconomic fluctuations. *Journal of Economic Perspectives*, *24*(4), 67–84. https://doi.org/10.1257/jep.24.4.67

Galavotti, I., Lippi, A., & Cerrato, D. (2021). The representativeness heuristic at work in decision-making: Building blocks and individual-level cognitive and behavioral factors. *Management Decision*, *59*(7), 1664–1683. https://doi.org/10.1108/MD-10-2019-1464

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, *103*(3), 592–596. https://doi.org/10.1037/0033-295X.103.3.592

Grether, D. M. (1992). Testing Bayes rule and the representativeness heuristic: Some experimental evidence. *Journal of Economic Behavior & Organization*, *17*(1), 31–57. https://doi.org/10.1016/0167-2681(92)90078-P

Gualtieri, S., & Denison, S. (2018). The development of the representativeness heuristic in young children. *Journal of Experimental Child Psychology*, *174*, 60–76. https://doi.org/10.1016/j.jecp.2018.05.006

Hamilton, K., Shih, S. I., & Mohammed, S. (2016). The development and validation of the rational and intuitive decision styles scale. *Journal of Personality Assessment*, *98*(5), 523–535. https://doi.org/10.1080/00223891.2015.1132426

Harren, V. A. (1979). A model of career decision making for college students. *Journal of Vocational Behavior*, *14*(2), 119–133. https://doi.org/10.1016/0001-8791(79)90065-4

Hodgkinson, G. P., & Clarke, I. (2007). Conceptual note: Exploring the cognitive significance of organizational strategizing: A dual-process framework and research agenda. *Human Relations*, *60*(1), 243–255. https://doi.org/10.1177/0018726707075297

Heath, C., Larrick, R. P., & Wu, G. (1999). Goals as reference points. *Cognitive psychology*, *38*(1), 79–109.

Hodgkinson, G. P., Sadler-Smith, E., Burke, L. A., Claxton, G., & Sparrow, P. R. (2009). Intuition in organizations: Implications for strategic management. *Long Range Planning*, *42*(3), 277–297. https://doi.org/10.1016/j.lrp.2009.05.003

Hong, C., & Feldman, G. (2023). *Revisiting the "belief in the law of small numbers": Conceptual replication and extensions Registered Report of problems reviewed in Tversky and Kahneman (1971)* [Stage 1: In-principle acceptance from Peer Community in Registered Report]. https://doi.org/10.17605/OSF.IO/MNS7J

Jané, M., Xiao, Q., Yeung, S., *Ben-Shachar, M. S., *Caldwell, A., *Cousineau, D., *Dunleavy, D. J., *Elsherif, M., *Johnson, B., *Moreau, D., *Riesthuis, P., *Röseler, L., *Steele, J., *Vieira, F., *Zloteanu, M., & ^Feldman, G. (2024). *Guide to Effect Sizes and Confidence Intervals*. http://dx.doi.org/10.17605/OSF.IO/D8C4G

Kahneman, D. (2002, December). Maps of bounded rationality: A perspective on intuitive judgment and choice. In *The Sveriges Riksbank Prize in economic sciences in memory of Alfred Nobel* (pp. 449–489). https://doi.org/10.1037/0003-066X.58.9.697

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge University Press. https://doi.org/10.1017/CBO9780511808098.004

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454. https://doi.org/10.1016/0010-0285(72)90016-3

Kassambara, A. (2020). *Ggpubr: "ggplot2" based publication ready plots* (Version 0.6.0). https://CRAN.R-project.org/package=ggpubr

Keren, G., & Teigen, K. H. (2004). Yet another look at the heuristics and biases approach. In D. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 89–109). John Wiley.

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Lakens, D., & Caldwell, A. R. (2021). Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science*, *4*(1), 251524592095150. https://doi.org/10.1177/2515245920951503

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science*, *1*, 389–402. https://doi.org/10.1177/2515245918787489

LeBel, E. P., Vanpaemel, W., Cheung, I., & Campbell, L. (2019). A brief guide to evaluate replications. *Meta-Psychology*, *3*, 1–9. https://doi.org/10.15626/MP.2018.843

Lee, M. D., & Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review*, *112*(3), 662–668. https://doi.org/10.1037/0033-295X.112.3.662

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge university press.

Lenth, R. V., Bolker, B., Buerkner, P., Giné-Vázquez, I., Herve, M., Jung, M., Love, J., Miguez, F., Riebl, H., & Singmann, H. (2023). *Emmeans: Estimated marginal means, aka least-squares means* (Version 1.9.0) [Computer software]. https://CRAN.R-project.org/package=emmeans

Li, M. Y., & Feldman, G. (2022). *Revisiting mental accounting classic paradigms: Replication of the problems reviewed in Thaler (1999)* [Stage 1 In-principle acceptance from Peer Community in Registered Report. Stage 2 preprint]. https://doi.org/10.17605/OSF.IO/V7FBJ

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433–442. https://doi.org/10.3758/s13428-016-0727-z

McCray, G. E., Purvis, R. L., & McCray, C. G. (2002). Project management under uncertainty: The impact of heuristics and biases. *Project Management Journal*, *33*(1), 49–57. https://doi.org/10.1177/875697280203300108

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes factors for common designs* (Version 0.8.12-4.6) [Computer software]. https://CRAN.R-project.org/package=BayesFactor

Norris, P., & Epstein, S. (2011). An experiential thinking style: Its facets and relations with objective and subjective criterion measures. *Journal of Personality*, *79*(5), 1043–1080. https://doi.org/10.1111/j.1467-6494.2011.00718.x

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*, 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Olson, C. L. (1976). Some apparent violations of the representativeness heuristic in human judgment. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(4), 599–608. https://psycnet.apa.org/doi/10.1037/0096-1523.2.4.599

Parsons, S., Kruijt, A. W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science*, *2*(4), 378–395.

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Revelle, W. (2024). *psych: Procedures for psychological, psychometric, and personality research* (Version 2.3.12) [Software]. Northwestern University. https://CRAN.R-project.org/package=psych

Röseler, L., Weber, L., Stich, E., Helgerth, K., Günther, M., Wagner, F.-S., & Schütz, A. (2022). *Measurements of susceptibility to anchoring are unreliable: Meta-analytic evidence from more than 50,000 anchored estimates*. https://doi.org/10.31234/osf.io/b6t35

Shiloh, S., Salton, E., & Sharabi, D. (2002). Individual differences in rational and intuitive thinking styles as predictors of heuristic responses and framing effects. *Personality and Individual Differences*, *32*(3), 415–429. https://doi.org/10.1016/S0191-8869(01)00034-4

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological science*, *26*(5), 559–569. https://doi.org/10.1177/0956797614567341

Smith, S. M., & Levin, I. P. (1996). Need for cognition and choice framing effects. *Journal of Behavioral Decision Making*, *9*(4), 283–290. https://doi.org/10.1002/(SICI)1099-0771(199612)9:4%3C283::AID-BDM241%3E3.0.CO;2-7

Stanovich, K. E., & West, R. F. (2000). Advancing the rationality debate. *Behavioral and Brain Sciences*, *23*(5), 701–717. https://doi.org/10.1017/S0140525X00623439

Stolwijk, S., & Vis, B. (2021). Politicians, the representativeness heuristic and decision-making biases. *Political Behavior*, *43*(4), 1411–1432. https://doi.org/10.1007/s11109-020-09594-6

Sundali, J., & Croson, R. (2006). Biases in casino betting: The hot hand and the gambler's fallacy. *Judgment and Decision Making*, *1*(1), 1–12. http://doi.org/10.1017/S1930297500000309

Teigen, K. H. (2022). Judgments by representativeness. In R. F. Pohl (Ed.), *Cognitive illusions: Intriguing phenomena in thinking, judgment, and memory* (pp. 191–208). Routledge. https://www.routledge.com/Cognitive-Illusions-Intriguing-Phenomena-in-Thinking-Judgment-and-Memory/Pohl/p/book/9780367724245

Thaler, R. H. (1999). Mental accounting matters. *Journal of Behavioral decision making*, *12*(3), 183–206. https://doi.org/10.1002/(SICI)1099-0771(199909)12:3%3C183::AID-BDM318%3E3.0.CO;2-F

Thaler, R. H. (2016). Behavioral economics: Past, present, and future. *American Economic Review*, *106*(7), 1577–1600. https://doi.org/10.1257/aer.106.7.1577

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105–110. https://doi.org/10.1037/h0031322

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, *185*(4157), 1124–1131.

Tversky, A., & Kahneman, D. (1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 84–98). Cambridge University Press.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315. https://psycnet.apa.org/doi/10.1037/0033-295X.90.4.293

Whelehan, D. F., Conlon, K. C., & Ridgway, P. F. (2020). Medicine and heuristics: Cognitive biases and medical decision-making. *Irish Journal of Medical Science*, *189*, 1477–1484. https://doi.org/10.1007/s11845-020-02235-1

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (Version 3.4.4). Springer. https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., & Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., Miller, E., & Smith, D. (2023). *haven: Import and export "SPSS," "Stata" and "SAS" files (Version 2.0.0) [Software]*. https://github.com/tidyverse/haven

Wilke, C. O. (2020). *cowplot: Streamlined plot theme and plot annotations for "ggplot2"* (Version 1.1.2). https://cran.r-project.org/web/packages/cowplot/index.html

Zhu, H. (2021). *kableExtra: Construct complex table with "Kable" and pipe syntax* (Version 1.3.4.9). https://rdrr.io/cran/kableExtra/

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*, Article e120. https://doi.org/10.1017/S0140525X17001972Sediti occus alitaquunt.

Revisiting representativeness classic paradigms:
Replication and extensions of nine experiments in Kahneman and Tversky (1972)

Supplementary

Contents

Open Science Disclosures

Procedure and Data Disclosures

Data collection

Data collection was completed before analyzing the data.

Conditions reporting

All collected conditions are reported.

Variables reporting

All variables collected for this study are reported and included in the provided data.

Overview of Experimental Design in the Target article

Original article methods

Type of study

Multiple study design (included both experimental manipulations and one-sample experiment).

Experimental design

The Sampling Distributions problem, the three Likelihood of Sampling Outcome problems, and the binomial Posterior Probabilities problem involved experimental manipulations.
The Sampling Distributions problem used a 3 x 3 between-subject design. One of the independent variables was the category of the sampling distributions, and the other independent variable was the sample size. The sampling distributions within the same category were then compared across different sample sizes. The original article did not mention if the participants were randomly assigned to the nine conditions.
The three Likelihood of Sampling Outcomes problems used a 1 x 2 between subject design. Participants were asked, in each of these problems, to determine if an outcome more/less extreme than the specified mean of probability would be more probable in a larger sample, smaller sample or about the same in both samples. The difference between the two conditions was a control for response bias, which was the question asking either if the critical value is above or below the specified mean. The numbers of participants who chose the larger sample, the smaller sample and the same in both were compared with one another.
The binomial Posterior Probabilities problem used a 2 x 5 between-subject design. One of the independent variables was the initial probability, which was defined by the proportion of cards marked as X and O in the deck. The other independent variable was the given sample proportion to the subject after drawing from a certain deck. The probabilities stated by the participants were then compared across different given sample proportions. The original article did not mention if the participants were randomly assigned to the nine conditions.

One-sample experiments

The two Sample-Population Similarity problems, Reflection of Randomness, and the non-binomial Posterior Probabilities problem did not use any experimental manipulations; they were all one-sample experiments.
In the Sample-Population problem ("birth sequence" scenario), the number of participants who reckoned the birth order of BGBBBB as more probable than GBGBBG was compared with the number of participants who reckoned otherwise. The number of participants who reckoned the birth order of BBBGGG as more probable than GBBGBG was compared with the number of participants who reckoned otherwise.
In the Sample-Population problem ("high school program" scenario), the number of participants who reckoned the sample class belonged to Program A was compared with the number of participants who reckoned the sample class belonged to Program B.

In the Reflection of Randomness problem, the number of participants who reckoned distribution A as more probable than distribution B was compared to the number of participants who reckoned otherwise.

For the non-binomial Posterior Probabilities problem, the odds reported by participants in scenario (i) was compared with the odds stated by the participants in scenario (ii).

Independent variables (IV)

In the Sample-Population Similarity problem, the IV was the similarity of sample to population (birth sequence in the first scenario, high school program in the second scenario). The IV was not manipulated.

In the Reflection of Randomness problem, the IV was the reflection of randomness in the sample, which was the distribution of marbles in this case. The IV was not manipulated.

In the Sample Distributions problem, the first IV was the category of the sampling distributions, and the second IV was the sample size. Both IVs were manipulated.

In the three Likelihood of Sampling Outcome problems, the IV in each problem was the level of extremeness of the outcome when compared to the specified mean of probability.

In the binomial Posterior Probabilities problem, the first IV was the initial probability, which was defined by the proportion of cards marked as X and O in the deck. The other IV was the given sample proportion to the subject after drawing from a certain deck. Both IVs were manipulated.

In the non-binomial Posterior Probabilities problem, the IV was different sample characteristics in the two different situations presented to participants. The IV was not manipulated.

Dependent variables (DV)

In the Sample-Population problem (birth sequence), the DV was the number of participants who reckoned the sample sequence to be more or less probable than the standard sequence.

In the Sample-Population Problem (high school program), the DV was the number of participants who reckoned the sample to be more likely to be from program A or program B.

In the Reflection of Randomness problem, the DV was the number of participants who reckoned type I or type II distribution to be more probable.

In the Sampling Distributions problem, the DV was the sampling distributions produced by the participants. The distributions were compared across different sample sizes to see if sample size was a factor the participants consider in making the sampling distribution.

In the three Likelihood of Sampling Outcome problems, the DV was the number of participants who reckoned the outcome more/less extreme than the specified mean of probability to be more probable in the smaller sample, bigger sample or the same in both.

In the binomial Posterior Probabilities problem, the DV was the probability stated by the participants about the sample being drawn from deck A.

In the non-binomial Posterior Probabilities problem, the DV was the odds stated by the participants about whether the sample consists of a certain characteristic.

Target article results

<u>Sample size before and after exclusions</u>

In the Sample-Population Similarity (birth sequence) problem, 92 participants were recruited and the original article did not mention whether if there were participants excluded.
In the Sample-Population Similarity (high school program) problem, 89 participants were recruited and the original article did not mention whether if there were participants excluded.
In the Reflection of Randomness problem, 52 participants were recruited and the original article did not mention whether if there were participants excluded.
In the Sampling Distributions problem, there were 9 conditions and the average sample size for each condition was 62, meaning that 558 participants were recruited. Tthe original article did not mention whether if there were participants excluded.
In the three Likelihood of Sampling Outcome problems, there were 97 participants recruited in total and they were divided approximately equally into the two conditions. In the babies scenario, there were 50 participants in the condition in which the outcome was more extreme than the specified mean of probability and 45 participants in the less extreme condition. In the investigator scenario, there were 49 participants in the condition in which the outcome was more extreme than the specified mean of probability and 48 participants in the less extreme condition. In the disease scenario, there were 48 participants in the condition in which the outcome was more extreme than the specified mean of probability and 48 participants in the less extreme condition. It appears that participants were excluded in in the babies scenario problem and the disease scenario problem, but the original article did not mention any exclusions.
In the binomial Posterior Probabilities problem, there were 10 conditions and the average sample size for each condition was 56, adding up to a total sample size of 560 participants. The original article did not mention whether any participants were excluded.
In the non-binomial Posterior Probabilities problem, 115 participants were recruited and the original article did not mention whether any participants were excluded.

<u>Sample description in the original</u>

For problem 1, 2, 4, 6 and 10, the participants students in grades 10,11 and 12 of college preparatory high schools with ages ranged from 15 to 18. No information related to their gender was provided. They were all from Israel and the survey was administered in quiz-like fashion by pen and paper in a natural classroom setting.
For problem 7, 8 and 9, the participants were Stanford undergraduates with no background in probability or statistics. No further information related to their age or gender was provided. The survey was also administered just like the problems above.
For problem 11, the participants were students from the University of Michigan and all of them had had at least one course in statistics before the survey. No further information related to their age or gender was provided. The survey was also administered just like the problems above.

Experimental design

For problem 6, there were nine conditions as there were three categories of sampling distributions and in each of them there were three different sample sizes. In each condition, the average number of participants in each group was 62. There was no statistical test conducted so no degree of freedom, p-value and effect size was available. The original article formatted a graph for each of the three categories of sampling distributions with the averaged sampling distributions from the participants. In each graph, the averaged sampling distributions of the three sample sizes were then compared. It was found that for all of the three categories of sampling distributions, the three averaged sampling distributions of the three different sample sizes had no differences. We conducted Kolmogorov-Smirnov tests between different subjective sampling distributions of different sample sizes in the same category based on the data given. In all of the comparisons, we found no support for the differences between any subjective sampling distributions of different sample sizes in the same category.

For problem 7, 8 and 9, these were a set of three questions asking whether an outcome more extreme than the specified mean of probability, was more probable in a larger sample, smaller sample or the same in both. There was another condition of this set of questions asking the same but about an outcome less extreme than the specified mean of probability. There were 97 participants in total and they were divided approximately equally into both conditions. Table 1 shows the counts of each option in each condition. The original article concluded that there was no difference among the three options within a condition in each of the problems. There was no statistical test conducted so no degree of freedom, *p*-value and effect size was available.

Table S1

Counts of Each Options Chosen by Participants in Each of the Problems in Both Conditions

	Condition of the outcome more extreme than the specified mean of probability			Condition of the outcome less extreme than the specified mean of probability		
	Larger sample	Smaller sample	The same in both sample	Larger sample	Smaller sample	The same in both sample
Problem 7	12	10*	28	9*	11	25
Problem 8	8	21*	20	10*	15	23
Problem 9	7	18*	23	14*	17	17

Note. The counts with * were the correct answers.

For problem 10, the participants were asked to estimate the posterior probability of a sample being drawn from a specified deck based on the sample proportion and the initial proportion of the population. There were two initial proportions of the population and in each of them, there were five different sample proportions, and so there were 10 conditions. There was an average of 56 participants in each condition. Table 2 showed the comparison of the mean subjective posterior probability by the participants in each condition. By median tests, it was found that in both populations, the estimates for 5:1 were significantly higher than those in the vertical column and those in the vertical column were significantly higher than the estimates for 18:14, *p<.01* in all comparisons. The original article did not provide the degrees of freedom and effect sizes.

Table S2

Comparison of Subjective Posterior Probability from Different Sample Proportions within the same population

Initial Proportion of population	p = 5/6			p = 2/3		
		4:2 .70			4:2 .68	
	18:14 .60	8:4 .70	5:1 .83	18:14 .58	8:4 .70	5:1 .85
		40:20 .70			40:20 .70	

Note. The upper entry of each cell is the sample presented; the lower entry is the median subjective estimate. The vertical column was the comparison of the median subjective estimates from the sample proportions with the same sample ratio. The horizontal row was the comparison of the median subjective estimates from the sample proportions with the same sample difference.

One sample experiment [no manipulation experiments]

For problem 1, there were 92 participants and they were asked to estimate the frequencies of three certain birth sequence given the frequency of a standard birth sequence. For the comparison between BGBBBB and GBGBBG the standard sequence, 75 participants reckoned BGBBBB to be less probable than the standard sequence and by a sign test, it was found that more participants more reckoned it to be less probable than the standard sequence, *p < .01*. For the comparison between BBBGGG and GBBGBG, no descriptive was available but by a sign test, more participants reckoned BBBGGG to be less probable than GBBGBG, *p < .01*. No degree of freedom or effect size was available. We conducted one-proportion z tests on the data given and found the same results.

For problem 2, there were 89 participants, and they were asked to determine if a class belongs to program A or B based on the information of gender proportion. 67 participants reckoned the class belongs to program A and by a sign test, it was found that more participants reckoned the class belongs to program A, *p < .01*. No degree of freedom or effect size was available. We conducted a one-proportion z test on the data given and found the same results.

For problem 4, there were 52 participants, and they were asked to determine which distribution of marbles among five people is more probable. 36 participants reckoned the nonuniform distribution to be more probable than the uniform distribution and by a sign test, it was found that more participants reckoned the nonuniform distribution to be more probable than the uniform distribution, *p < .01*. No degree of freedom or effect size was available. We conducted a one-proportion z test on the data given and found the same results.

For problem 11, there were 115 participants, and they were asked to determine the two odds of two samples, (i) and (ii), drawn from a population that consists of certain characteristics. The median subjective odds were 8 in case (i) and 2.5 in case (ii), and by a median test, it was found that the odds of case (i) stated by the participants were larger than the one in case (ii), $p < .01$. No degree of freedom or effect size was available.

Effect size calculations of the target article effects

Effect sizes were not reported in the original article. We thus computed these whenever possible (when there was sufficient information). The R code used to calculate the effect size can be accessed using the OSF link provided in the manuscript.

Sample-Population Similarity (birth sequence) problem

The original article conducted a sign test, which is a test of whether a certain proportion is significantly larger or smaller than an expected proportion in a population. It is similar to a binomial test. We used Cohen's *h* to quantify the effect (Cohen, 1988). For the first problem (birth sequence scenario), 75 out of 92 participants reckoned the birth sequence BGBBBB as less probable than the birth sequence GBGBBG (the expected proportion by chance is 0.5). Cohen's *h* is 0.68. For the other comparison, between birth sequence BBBGGG and GBBGBG, we could not calculate an effect size as the original article did not provide the number of responses to the choice options. Regardless, this was not a comparison of interest in the original article.

Sample-Population Similarity (high-school programs) problem

67 out of 89 participants reckoned the class to be program A (expected proportion is 0.5). Cohen's *h* is 0.53.
Reflection of Randomness problem
The original article conducted a sign test. 36 out of 52 participants reckoned the nonuniform distribution to be more probable than the uniform distribution and the expected proportion was also 0.5 by chance. Cohen's *h* is 0.39.

Sampling Distributions problem

The original article did not conduct any statistical tests but simply showed the comparisons in graphs. As this problem concerned a comparison between distributions, we used two-sample Kolmogorov-Smirnov tests. KT hypothesized that the subjective sampling distribution would not differ between different sample size conditions. The effect size for the two-sample Kolmogorov-Smirnov test is *D*.
For distribution of sexes, the comparison between sampling distributions with $N = 10$ and $N = 100$ had an effect size *D* of 0.09.
For distribution of sexes, the comparison between sampling distributions with $N = 100$ and $N = 1000$ had an effect size *D* of 0.18.
For distribution of sexes, the comparison between sampling distributions with $N = 10$ and $N = 1000$ had an effect size *D* of 0.18.
For distribution of heartbeat type, the comparison between sampling distributions with $N = 10$ and $N = 100$ had an effect size *D* of 0.18.

For distribution of heartbeat type, the comparison between sampling distributions with *N* = 100 and *N* = 1000 had an effect size *D* of 0.18.
For distribution of heartbeat type, the comparison between sampling distributions with *N* = 10 and *N* = 1000 had an effect size *D* of 0.09.
For distribution of height, the comparison between sampling distributions with N = 10 and *N* = 100 had an effect size *D* of 0.
For distribution of height, the comparison between sampling distributions with *N* = 100 and *N* = 1000 had an effect size *D* of 0.29.
For distribution of height, the comparison between sampling distributions with *N* = 10 and *N* = 1000 had an effect size *D* of 0.29.

Likelihood of Sampling Outcomes problems

The original article did not conduct any statistical tests. As the original article hypothesized that participants would show no preference for the correct answer, we tested to whether the proportion of participants who chose the correct answer was larger than the expected proportion. We conducted one proportion *z*-tests were and used Cohen’s *h* as the effect size.

Babies scenario.

Cohen’s *h* in the “more extreme” condition = -0.30. Cohen’s *h* in the “less extreme” condition = -0.30.

Investigator scenario.

Cohen’s *h* in the “more extreme” condition = 0.20. Cohen’s *h* in the “less extreme” condition = -0.28.

Disease scenario.

Cohen’s *h* in the “more extreme” condition = 0.08. Cohen’s *h* in the “less extreme condition” = -0.09.

Posterior Probabilities (binomial) problem

Median tests were conducted to test whether the median subjective estimate of 5:1 was larger than the ones of 4:2, 8:4, and 40:20. Median tests were also conducted to test whether the median subjective estimates of 18:14 was smaller than the ones of 4:2, 8:4 and 40:20. We could not compute effect sizes in the original article due to insufficient information.
Posterior Probabilities (non-binomial) problem
A median test was conducted to compare the odds reported by participants in case (i) with the odds in case (ii). However, like the previous problem, we could not compute effect sizes in the original article due to insufficient information.

Power analysis of target article effects

We conducted a power analysis using the effect sizes that we calculated in the target article. We aimed for .95 power, using a standard 5% alpha error rate. The largest required sample size among all of the problems is 334. As we randomly assigned participants to receive five out nine problems, we doubled the sample size, resulting in a target sample size of 668 participants.
See additional analyses in the accompanying Rmarkdown.

Sample-Population Similarity

“Birth sequence” scenario.

28 participants required to detect Cohen’s *h* 0.68.

“High school program” scenario.

47 participants required to detect Cohen’s *h* 0.53.

Reflection of Randomness

84 participants required to detect Cohen’s *h* 0.39.

Sampling Distributions

We did not conduct a power analysis for this problem as we could not find any accessible methods to conduct power analysis for Kolmogorov-Smirnov tests.
Likelihood of Sampling Outcome
Note that these problems tested a null hypothesis. Thus, the effect sizes were small, which require very large sample sizes.

“Babies” scenario (“more extreme” condition).

141 participants required to detect Cohen’s *h* -0.30.

“Babies” scenario (“less extreme” condition).

141 participants required to detect Cohen’s *h* -0.30.

“Investigator” scenario (“more extreme condition”).

337 participants required to detect Cohen’s *h* 0.20.

“Investigator” scenario (“less extreme condition”).

162 participants required to detect Cohen’s *h* -0.28.

“Disease” scenario (“more extreme condition”).

[Null hypothesis]
1,1711 participants required to detect Cohen’s *h* 0.09.

“Disease” scenario (“less extreme condition”).

[Null hypothesis]
1,606 participants required to detect Cohen’s *h* -0.09.

The power analysis for problems that tested a null hypothesis should be based on the smallest effect size of interest (SESOI). We determined the SESOI based on the effect sizes of the previous problems in the original. We took the smallest effect size (Cohen’s *h*=0.39, in Problem 4) and divided it by half. It was then used in a power analysis to determine the required sample size for problem 7, 8 and 9, which resulted in an estimated sample size of 334.
The R code is provided below:

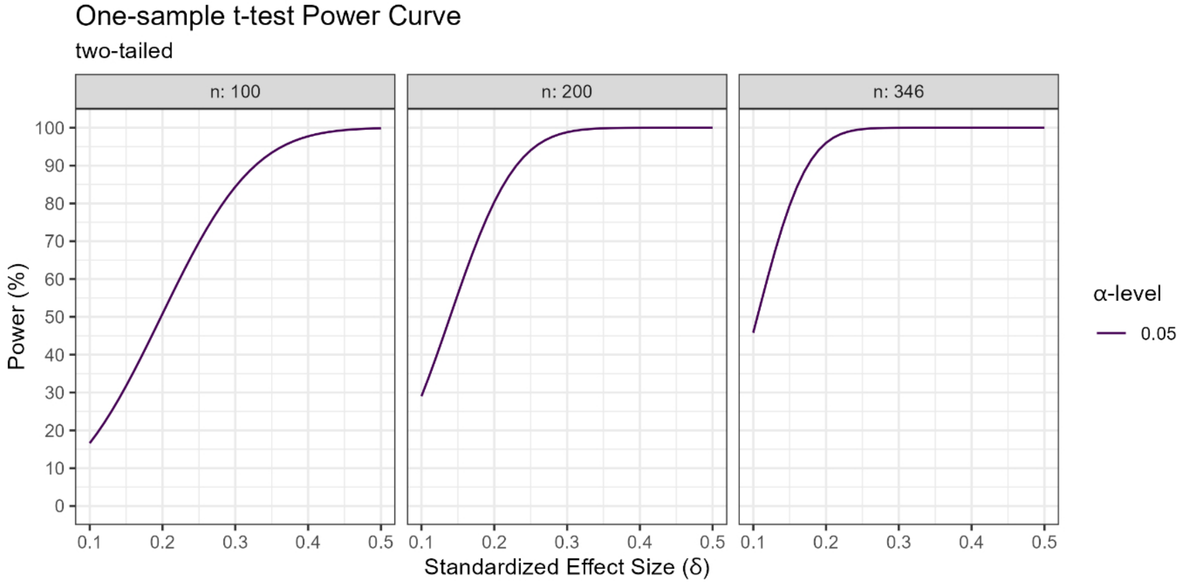
```
power.proportions (h=0.3947911/2, power=0.95, sig.level=0.05, type="one")

    proportion power calculation for binomial distribution (arcsine transformation)

              n = 333.4969
          power = 0.95
              h = 0.1973956
      sig.level = 0.05

NOTE: n is the number of observations
```

For Problems 10 and 11, as previously mentioned, the effect size could not be calculated as the details of the data were not provided in the original article. Therefore, the calculation of the required sample size was based on the previous problems.

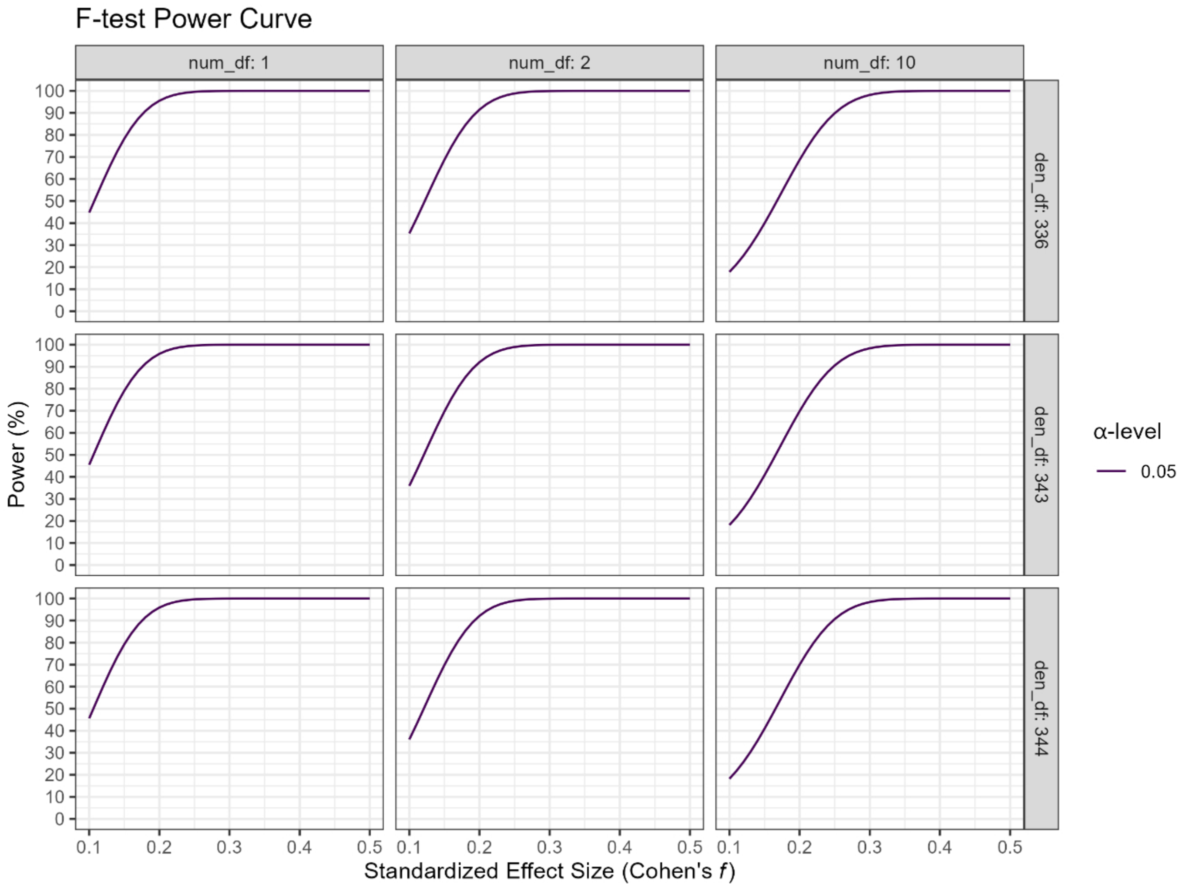
Sensitivity analyses on final sample

Final sample of 623. Five out of nine problems, therefore on average 346 per problem.
See code in the accompanying Rmarkdown. We plotted power curves for each test for a range of sample sizes (maximum sample size is the final sample in the replication) using the SuperPower R package (Lakens & Caldwell, 2021). Note that we did could not find a method to conduct a sensitivity analysis for Problem 6 which used Kolmogorov–Smirnov test.

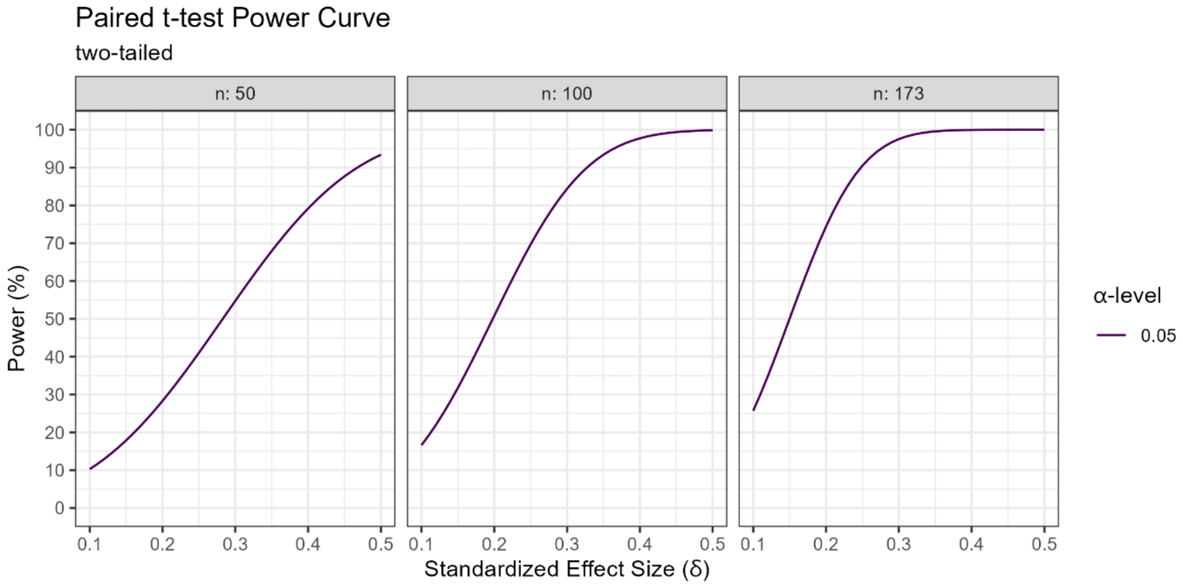
Problems 1, 2, 4, and 7-9

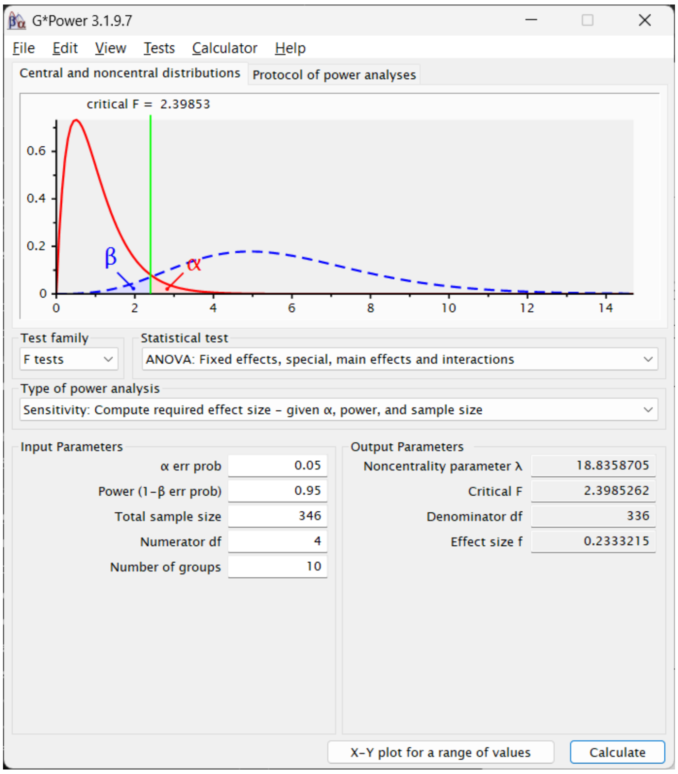


Problem 10 (ANOVA with 10 conditions)



Problem 11 (paired-sample *t*-test)




G*Power 3.1.9.7
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
critical F = 2.39853
0.6
0.4
0.2
0
0
2
4
6
8
10
12
14
β
α
X-Y plot for a range of values
Calculate

Test family: F tests

Statistical test: ANOVA: Fixed effects, special, main effects and interactions

Type of power analysis: Sensitivity: Compute required effect size – given α, power, and sample size

Input Parameters

α err prob: 0.05
Power (1–β err prob): 0.95
Total sample size: 346
Numerator df: 4
Number of groups: 10

Output Parameters

Noncentrality parameter λ: 18.8358705
Critical F: 2.3985262
Denominator df: 336
Effect size f: 0.2333215

Decision Styles Scale

Table S3

Decision Styles Scale (DSS): List of statements

Decision Styles	Statements
Rational Decision Style	1. I prefer to gather all the necessary information before committing to a decision.
	2. I thoroughly evaluate decision alternatives before making a final choice.
	3. In decision making, I take time to contemplate the pros/cons or risks/benefits of a situation.
	4. Investigating the facts is an important part of my decision-making process.
	5. I weigh a number of different factors when making decisions.
Intuitive Decision Style	1. When making decisions, I rely mainly on my gut feelings.
	2. My initial hunch about decisions is generally what I follow.
	3. I make decisions based on intuition.
	4. I rely on my first impressions when making decisions.
	5. I weigh feelings more than analysis in making decisions.

Note. Instructions read: “The following questions relate to how you make decisions. There are no "right" or "wrong" answers, so please state your opinion as honestly as possible. Using the scale below, please indicate the extent to which you agree or disagree with the statements. Describe how you are now, not as you wish to be in the future.

Overview of the Problems Included in the Replication

Table S4

Replication: Problems, design, and predictions

#	Domain	Design	Problem	Predictions	Correct answer
1	Similarity of sample to population (birth sequence)	No manipulation; one sample	All families of six children in a city were surveyed. In 72 families the exact order of births of boys and girls was G B G B B G. What is your estimate of the number of families surveyed in which the exact order of births was B G B B B B / B B B G G G / G B B G B G?	1a: Sample with boy-girl split closer to expected equal 50-50 split in the population (GBGBBG) is perceived as more probable than a lesser equal split sequence (BGBBBB) 1b: Sample with less orderly sequence (GBBGBG) is perceived as more probable than a sample with an orderly sequence (BBBGGG)	"The two birth sequences are about equally likely" (p. 432) 1a: equal probability 1b: equal probability.
2	Similarity of sample to population (gender proportion)	No manipulation; one sample	There are two programs in a high school. Boys are a majority (65%) in program A, and a minority (45%) in program B. There is an equal number of classes in each of the two programs. You enter a class at random, and observe that 55% of the students are boys. What is your best guess - does the class belong to program A or to program B?"	2: When observing a class with 55% boys, class is perceived to be more likely Program A (65% boys) than Program B (45% boys) given that boys are a majority and therefore more "representative".	"In fact, it is slightly more likely that the class belongs to program B (since the variance for p = .45 exceeds that for p = .65)." (p. 433)

#	Domain	Design	Problem	Predictions	Correct answer
4	Reflection of randomness in the sample	No manipulation; one sample	On each round of a game, 20 marbles are distributed at random among five children: Alan, Ben, Carl, Dan, and Ed. Consider the following distributions. Type I Type II Alan 4 4 Ben 4 4 Carl 5 4 Dan 4 4 Ed 3 4 In many rounds of the game, will there be more results of type I or of type II?	Type II distribution is perceived as more probable than Type II distribution.	“The uniform distribution of marbles (II) is, objectively, more probable than the nonuniform distribution (I)” (p. 434)
6a	Sampling Distributions: Distribution of Sexes (Binomial, p = .50)	3 conditions between-subject (sample size): *N* = 10, *N* = 100, *N* = 1000	[10/100/1000] babies are born everyday in a certain region. Given that the possibilities of both gender are equal (50/50), on what percentage of days will the number of boys among [10/100/1000] babies be as follows: (Note that the categories include all possibilities, so your answers should add up to about 100%). __ [0 boys/Up to 5 boys/Up to 50 boys] (1) __ [1 boy/5 to 15 boys/50 to 150 boys] (2) __ [2 boys/15 to 25 boys/150 to 250 boys (3) __ [3 boys/25 to 35 boys/250 to 350 boys (4) __ [4 boys/35 to 45 boys/350 to 450 boys (5) __ [5 boys/45 to 55 boys/450 to 550 boys (6) __ [6 boys/55 to 65 boys/550 to 650 boys (7) __ [7 boys/65 to 75 boys/650 to 750 boys (8) __ [8 boys/75 to 85 boys/750 to 850 boys (9) __ [9 boys/85 to 95 boys/850 to 950 boys (10) __ [10 boys/More than 95 boys/More than 950 boys (11) Note: The means of estimate of each row of each subject were taken to make the mean sampling distributions.	[KT’s null effect hypothesis] Law of small numbers / Sample size neglect: There would be no differences in distribution comparing condition with 10, 100, or 1000. [Competing, reframed from the null effect] Law of big numbers / Sample size sensitivity There would be differences in distribution comparing condition with 10, 100, or 1000.	

#	Domain	Design	Problem	Predictions	Correct answer
6b	Sampling Distributions: Distribution of Heartbeat Type (Binomial, p = .80)	3 conditions between-subject (sample size): *N* = 10, *N* = 100, *N* = 1000	[10/100/1000] babies are born everyday in a certain region. Given that 80% of all newborns have a heartbeat of type α and the remaining 20% have a heartbeat of type β, on what percentage of days will the number of babies with heartbeat of type α among [10/100/1000] be as follows (Note that the categories include all possibilities, so your answers should add up to about 100%). __ [0 babies/Up to 5 babies/Up to 50 babies] (1) __ [1 baby/5 to 15 babies/50 to 150 babies] (2) __ [2 babies/15 to 25 babies/150 to 250 babies (3) __ [3 babies/25 to 35 babies/250 to 350 babies (4) __ [4 babies/35 to 45 babies/350 to 450 babies (5) __ [5 babies/45 to 55 babies/450 to 550 babies (6) __ [6 babies/55 to 65 babies/550 to 650 babies (7) __ [7 babies/65 to 75 babies/650 to 750 babies (8) __ [8 babies/75 to 85 babies/750 to 850 babies (9) __ [9 babies/85 to 95 babies/850 to 950 babies (10) __ [10 babies/More than 95 babies/More than 950 babies (11) Note: The means of estimate of each row of each subject were taken to make the mean sampling distributions.	[KT's null effect hypothesis] Law of small numbers / Sample size neglect: There would be no differences in distribution comparing condition with 10, 100, or 1000. [Competing, reframed from the null effect] Law of big numbers / Sample size sensitivity There would be differences in distribution comparing condition with 10, 100, or 1000.	
6c	Sampling Distributions: Distribution of height.	3 conditions between-subject (sample size): *N* = 10, *N* = 100, *N* = 1000	A regional induction centre records the average height of the [10/100/1000] men who are examined every day. Given that the average height of the male population lies between 170-175cm and the frequency of heights decreases with the distance from the mean, on what percentage of men's different height classes will be recorded on a certain day as follows: __ Up to 160cm (1) __ 160-165cm (2) __ 165-170cm (3) __ 170-175cm (4) __ 175-180cm (5) __ 180-185cm (6) __ More than 185cm (7) (Note that the categories include all possibilities, so your answers should add up to about 100%)	[KT's null effect hypothesis] Law of small numbers / Sample size neglect: There would be no differences in distribution comparing condition with 10, 100, or 1000. [Competing, reframed from the null effect] Law of big numbers / Sample size sensitivity There would be differences in distribution comparing condition with 10, 100, or 1000.	

#	Domain	Design	Problem	Predictions	Correct answer
7	Likelihood of Sampling Outcomes in Small vs. Large Samples Size of hospital	2 conditions between-subject (more versus less)	A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower. For a period of 1 year, each hospital recorded the days on which [more/less] than 60% of the babies born were boys. Which hospital do you think recorded more such days? (The larger hospital/The smaller hospital/About the same (i.e., within 5% of each other.	People tend to judge the two hospitals as having the same likelihood for 60% boys.	Smaller hospital has larger variance and therefore more likely to have a day with 60%.
8	Likelihood of Sampling Outcomes in Small vs. Large Samples Line vs. page	2 conditions between-subject (more versus less)	An investigator studying some properties of language selected a paperback and computed the average word-length in every page of the book (i.e., the number of letters in that page divided by the number of words). Another investigator took the first line in each page and computed the line's average word-length. The average word-length in the entire book is 4. However, not every line or page has exactly that average. Some may have a higher average word-length, some lower. The first investigator counted the number of pages that had an average word-length of 6 or [more/less] and the second investigator counted the number of lines that had an average word-length of 6 or [more/less]. Which investigator do you think recorded a larger number of such units (pages for one, lines for the other)? (The page investigator; The line investigator; About the same (i.e., within 5% of each other))	People tend to judge the two investigators as having the same likelihood of having an average of 6 or more words per unit.	Line has smaller sample and larger variance and therefore more likely to have average word-length of 6 or more than page.

#	Domain	Design	Problem	Predictions	Correct answer
9	Likelihood of Sampling Outcomes in Small vs. Large Samples 3 men versus 1 man	2 conditions between-subject (more versus less)	A medical survey is being held to study some factors pertaining to coronary diseases. Two teams are collecting data. One checks 3 men a day, and the other checks 1 man a day. These men are chosen randomly from the population. Each man's height is measured during the checkup. The average height of adult males is 5 ft 10 in., and there are as many men whose weight is above average as there are men whose height is below average. The team checking 3 men a day ranks them with respect to their height, and count the days on which the height of the middle man is [more/less] than 5 ft 11 in. The other team checking 1 man a day merely counts the days on which the man they checked was [taller/shorter] than 5 ft 11 in. Which team do you think counted more such days? The team checking 3 men; The team checking 1 man; About the same (i.e., within 5% of each other)	People tend to judge the medical surveys as having the same likelihood of men taller than 5 ft 10 in.	1 man a day is smaller and has larger variance than 3 men a day, and therefore more likely to record
10	Posterior Probabilities Binomial task	2 x 5 between-participants design	Consider two very large decks of cards, denoted A and B. In deck A, [5/6; 2/3; 5/6; 2/3; 5/6; 2/3; 5/6; 2/3; 5/6; 2/3] of the cards are marked X and [1/6; 1/3; 1/6; 1/3; 1/6; 1/3; 1/6; 1/3; 1/6; 1/3] are marked O. In deck B, [1/6; 1/3; 1/6; 1/3; 1/6; 1/3; 1/6; 1/3; 1/6; 1/3] of the cards are marked X, and [5/6; 2/3; 5/6; 2/3; 5/6; 2/3; 5/6; 2/3; 5/6; 2/3] are marked O. One of the decks has been selected by chance, and [12; 12; 6; 6; 60; 60; 6; 6; 32; 32] cards have been drawn at random from it, of which [8; 8; 4; 4; 40; 40; 5; 5; 18; 18] are marked X and [4; 4; 2; 2; 20; 20; 1; 1; 14; 14] are marked O. What do you think the probability is that the [12; 12; 6; 6; 60; 60; 6; 6; 32; 32] cards were drawn from deck A, that is from the deck in which most of the cards are marked X? For example, if you think that there is a 100% chance that the sample was drawn from deck A, you can input "1". If you think that there is a 60% chance that the sample was drawn from deck A, you can input "0.6".	People tend to rely on sample proportions of the two objects (as this is the most representative feature). In both pairs of population proportions (5/6 and 1/6 vs. 2/3 and 1/3), participants' posterior estimates in the 5:1 sample proportion condition would be larger than in the 4:2, 8:4, and 40:20 conditions, which would be larger than in the 18:14 conditions.	"In the symmetric binomial task the objective posterior probability depends only on the difference between the numbers of red and blue chips observed in the sample. posterior odds are given by (p/l-p)^(r-b)" (p. 446-8)

#	Domain	Design	Problem	Predictions	Correct answer
11	Posterior Probabilities Non-Binomial Task	2 conditions within-participants (single person vs. 6 persons)	The average heights of adult males and females in the US are, respectively, 5 ft 10 in. and 5 ft 4 in. Both distributions are approximately normal with a standard deviation of about 2.5 in. An investigator has selected one population by chance and has drawn from it a random sample. (i) What do you think is the probability in percentage that he has selected the male population if the sample consists of a single person whose height is 5 ft 10 in.? (ii) What do you think is the probability in percentage that he has selected the male population if the sample consists of 6 persons whose average height is 5 ft 8 in.?	In a population with average heights of 5 ft 10 in. for males and 5 ft 4 for females, people tend to perceive a randomly drawn single person with 5 ft 10 in. as more likely to drawn from a male population than randomly drawn 6 persons averaging 5 ft 8 in.	"The correct odds are 16% in case (i) and 29% in case (ii)." (p. 449)

Deviations from Preregistration

Table S5

Pre-registration plan versus final report

Components in preregistration	**Location of preregistered decision/plan**	**Deviations**	**Description of deviation**	**Rationale for deviation**	**Impact of deviation on results**	**Date/time of decision for deviation + stage**
Study design	p. 26-30, Method, "Design and Procedure", link: https://osf.io/nbdjr/	no	/	/	/	/
Measured variables	p. 30-31, Method, "Measures", Link: https://osf.io/nbdjr/	no	/	/	/	/
Exclusion criteria	p. 22-23, Generalized exclusion criteria and Specific criteria, "Exclusion criteria', Link: https://osf.io/ge9n4/	minor	The pre-registration stated that the analysis would focus on the full sample while the final report's analysis focused on the excluded sample.	The survey of this replication is very cognitively demanding for the participants and so the exclusion criteria is crucial to maintain the reliability of the data.	The full sample analysis is available in "Full Sample Analysis (No Exclusions)" under "Additional analyses and results" in this supplementary. No major difference was spotted.	25-26 May 2020, after pre-registration but before data collection
IV	p. 32-35, Method, "Table 6", Link: https://osf.io/nbdjr/	no	/	/	/	/
DV	p. 32-35, Method, "Table 6", Link: https://osf.io/nbdjr/	no	/	/	/	/
Data analysis	p.35-38, Method, "Evaluation criteria for replication findings" and "Replication evaluation", link: https://osf.io/nbdjr/	no	/	/	/	/

Full Sample Results (Without Exclusions)

Table 1

Sample Comparison Between the target article and the Replication

	Kahneman and Tversky (1972)	Replication sample
Sample size	Approximately 1500 in total (different participants responded to different problems, with some responding to 2-4 problems)	623
Type of sample	High-school students (Problems 1-4 and 8) Undergraduates (Problems 5-7 and 11-12)	MTurk workers on CloudResearch
Geographic origin	Israel (Problems 1-4 and 8) US (Problem 5-7 and Problems 11-12)	US American
Gender	Not specified	352 males, 327 females, 4 other/would rather not disclose
Median age (years)	Not specified	40
Average age (years)	Not specified	42
Age range (years)	15-18 (Israeli high school students), not specified (other samples)	21-91
Medium (location)	Pen and paper in a classroom situation	Computer (online)
Compensation	Not specified	Nominal payment
Year	Not specified	2020

Table 2

Summary of target article's findings

Problem	**Factors**		***p***	**Effect [95% CI]**
1. Samp-population similarity (birth sequence)	/		<.001	Cohen's *h*=0.68 [0.48, 0.89]
2. Samp-population similarity (high-school prog.)	/		<.001	Cohen's *h*=0.53 [0.32, 0.74]
4. Reflection of randomness	/		.007	Cohen's *h*=0.39 [0.12, 0.67]
6. Sampling dist.	Gender distribution	N=10 vs N=100	1.00	
		N=100 vs N=1000	.993	
		N=10 vs N=1000	.993	
	Heartbeat distribution	N=10 vs N=100	.993	
		N=100 vs N=1000	.993	
		N=10 vs N=1000	.993	
	Height distribution	N=10 vs N=100	1.00	
		N=100 vs N=1000	.938	
		N=10 vs N=1000	.938	
7. Likelihood of Sampling Outcomes (babies)	"More extreme" condition [a]		.968	Cohen's *h*=-0.30 [-0.58, -0.02]
	"Less extreme" condition		.959	Cohen's *h*=-0.30 [-0.60, -0.01]
8. Likelihood of Sampling Outcomes (investigator)	"More extreme" condition		.103	Cohen's *h*=0.20 [-0.08, 0.48]
	"Less extreme" condition		.954	Cohen's *h*=-0.28 [-0.57, 0.00]
9. Likelihood of Sampling Outcomes (disease)	"More extreme" condition		.323	Cohen's *h*=0.08 [-0.20, 0.37]
	"Less extreme" condition		.677	Cohen's *h*=-0.09 [-0.37, 0.19]
10. Posterior probability (binomial)	Initial proportion: 5:1	5:1 vs 4:2	<.01	
		5:1 vs 8:4	<.01	
		5:1 vs 40:20	<.01	
		18:14 vs 4:2	<.01	
		18:14 vs 8:4	<.01	
		18:14 vs 40:20	<.01	
	Initial proportion: 2:1	5:1 vs 4:2	<.01	
		5:1 vs 8:4	<.01	
		5:1 vs 40:20	<.01	
		18:14 vs 4:2	<.01	
		18:14 vs 8:4	<.01	
		8:14 vs 40:20	<.01	
11. Posterior probability (non-binomial)	/		<.01	

Note. Problem 1 included two questions but effect size could only be calculated for the first question. Problems 6, 7, 8, and 9 tested null hypotheses. Therefore, *p*-values were large and effect sizes were small, and reflect a one-tail t-test of the directionality of the prediction (which is why confidence intervals might not include the null, yet have very high p-values).

[a] More extreme condition = Outcome more extreme than the specified mean of probability, Less extreme condition = Outcome less extreme than the specified mean of probability.

Table 5

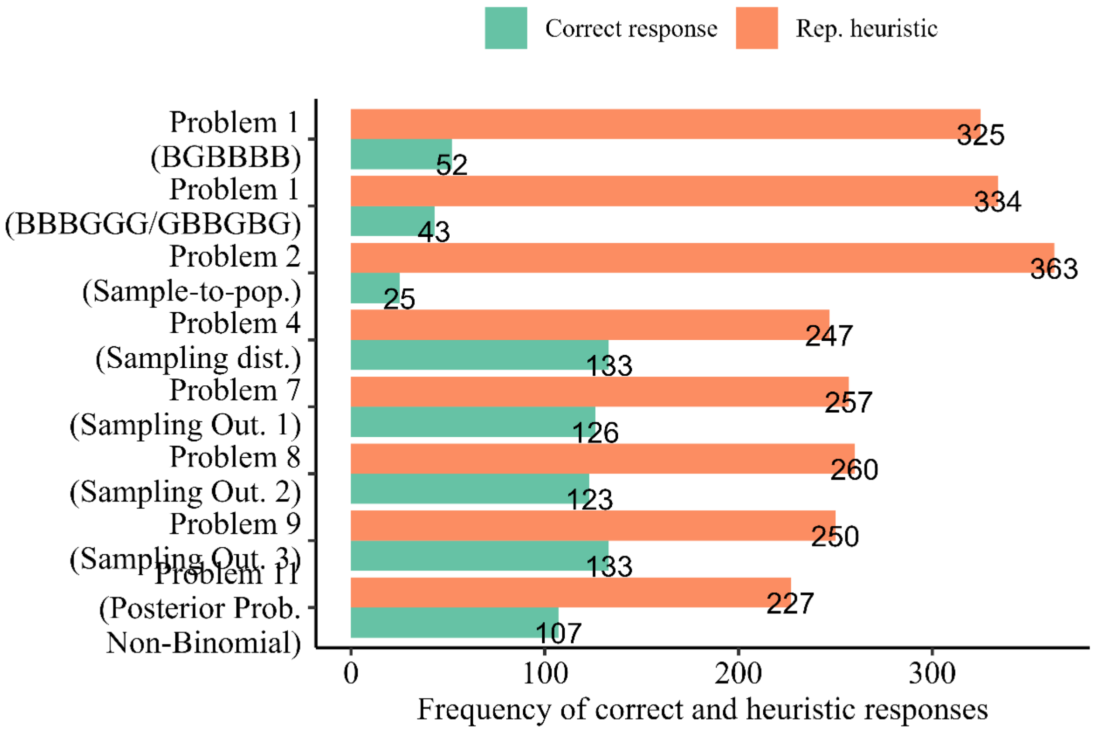
Replication: Descriptive Statistics for Problems That Scored the Representativeness Heuristic

Problems		**Option**	**Count**	***N***
1: Sample-to-population similarity (birth sequence)	Birth sequence BGBBBB	Less than 72	325	377
		Equal or more than 72*	52	
	Birth sequence BBBGGG vs GBBGBG	BBBGGG equal or more probable*	43	377
		GBBGBG more probable	334	
2: Sample-to-population similarity (high-school program)		Program A	363	388
		Program B*	25	
4: Sampling distributions		Distribution I (non-uniform)	247	380
		Distribution II (uniform)*	133	
7-9: Likelihood of sampling outcomes				
7: *Babies*	More extreme	About the same	75	192
		The smaller hospital*	59	
		The larger hospital	58	
	Less Extreme	About the same	74	191
		The smaller hospital	50	
		The larger hospital*	67	
8: *Investigator*	More extreme	About the same	68	
		The line investigator*	79	192
		The page investigator	45	
	Less Extreme	About the same	72	191
		The line investigator	75	
		The page investigator*	44	
9: *Disease*	More extreme	About the same	55	192
		The team checking 1*	39	
		The team checking 3	98	
	Less extreme	About the same	59	191
		The team checking 1	38	
		The team checking 3*	94	

Note. Correct answers (no use of representativeness heuristic) are starred.

Figure 1

Frequency of Heuristic Responses



We summarized the findings for Problems 1, 2, and 4 in Table 6.

Table 6

Problems 1 and 2 (Sample-to-Population Similarity) and Problem 4 (Reflection of Randomness): Comparison of the findings in target article versus replication

		Target article		**Replication**		
Problem		***p***	**Cohen's *h*** **[95% CI]**	***p***	**Cohen's *h*** **[95% CI]**	**Interpretation**
1. Sample-Population Similarity (birth sequences)	BGBBBB vs GBGBBG	< .001	0.68 [0.48, 0.89]	< .001	0.81 [0.71, 0.91]	Signal–consistent
	BBBGGG vs GBBGBG	< .01	/	< .001	0.88 [0.78, 0.98]	Signal NA (effect size of the target article is not available)
2. Sample-Population Similarity (high school programs)		< .001	0.53 [0.32, 0.74]	< .001	1.06 [0.96, 1.16]	Signal–inconsistent, larger
4. Reflection of Randomness		.007	0.39 [0.12, 0.67]	< .001	0.30 [0.20, 0.41]	Signal–consistent

Note. Sign tests were conducted in the target article and one-proportion z-tests in the replication.

Problem 1 (Sample-to-Population Similarity, Birth Sequence)

Consistent with the target article, most participants selected the heuristic response. In Problem 1 (birth sequence), one-proportion z-tests indicated that most participants (325 out of 377) estimated the birth sequence BGBBBB to be less probable than the birth sequence GBGBBG, $\chi^2 = 196$, $p < .001$, $h = 0.81$, 95% CI [0.71, 0.91]. Most participants (334 out of 377) estimated the birth sequence BBBGGG to be less probable than the birth sequence GBBGBG, $\chi^2 = 223$, $p < .001$, $h = 0.88$, 95% CI [0.78, 0.98].

Problem 2 (Sample-to-Population Similarity, High-School Program)

In Problem 2 (high school program), we conducted a one-proportion z-test and found that most participants (363 out of 388 participants) estimated that the class belonged to program A rather than program B, $\chi^2 = 293$, $p < .001$, $h = 1.06$, 95% CI [0.96, 1.16]. We concluded that our findings are consistent with the target article's.

Problem 4 (Reflection of Randomness)

We conducted a one-proportion *z*-test and found that most participants (247 out of 380 participants) estimated distribution I (the non-uniform distribution) to be more probable than distribution II (the uniform distribution), $\chi^2 = 34$, $p < .001$.

Problem 6 (Sampling Distributions)

We summarized the comparison of the statistical details between the replication and the original findings in Table 7. We plotted participants' mean probability estimates in the three scenarios in Figure 2. All three distributions were consistent with the target article's findings.

Table 7

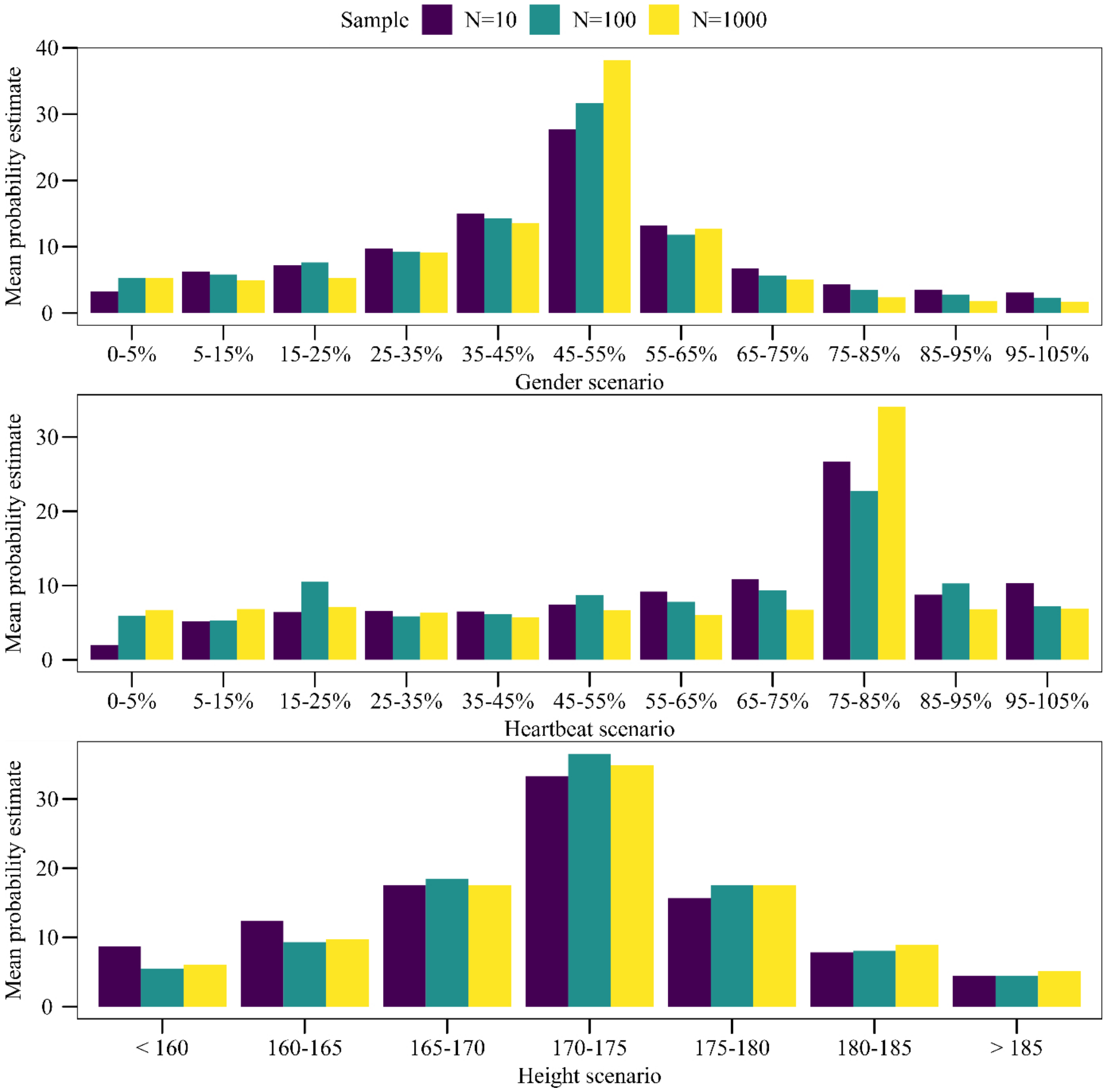
Problem 6 (Sampling Distributions): Comparison of findings in target article versus replication

		Target article	**Replication**	
Categories of sampling distributions	**Comparisons of sampling distributions**	***p***	***D***	***p***
Distribution of genders	$N = 10$ vs $N = 100$	1.00	0.18	1.00
	$N = 100$ vs $N = 1000$	.993	0.36	.480
	$N = 10$ vs $N = 1000$	.993	0.27	.830
Distribution of blood type	$N = 10$ vs $N = 100$	.993	0.18	1.00
	$N = 100$ vs $N = 1000$	.993	0.54	.075
	$N = 10$ vs $N = 1000$	.993	0.46	.210
Distribution of height	$N = 10$ vs $N = 100$	1.00	0.29	.960
	$N = 100$ vs $N = 1000$	.938	0.29	.960
	$N = 10$ vs $N = 1000$	.938	0.14	1.00

Note. We conducted Kolmogorov-Smirnov tests on the given data of the target article. Kolmogorov-Smirnov tests were also conducted in the replication. *D* is the effect size for Kolmogorov-Smirnov tests.

Figure 2

Problem 6 (Sampling Distributions): Mean Probability Estimates of Sampling Distributions



The target article did not conduct any statistical tests for this problem. We conducted a Kolmogorov-Smirnov test on each comparison of sample size ($N = 10$ vs $N = 100$, $N = 100$ vs $N = 1000$, and $N = 10$ vs $N = 1000$) in each category of the sampling distribution (distribution of gender, blood type, and height). We did not find evidence for differences in mean probability estimates between sample size conditions in any of the categories. These results are consistent with the target article. We could not quantify the null as we found no Bayesian approach for Kolmogrorov-Smirnov tests.

Problems 7-9 (Likelihood of Sampling Outcomes)

We summarized the comparison of the statistical details between the target article and the replication for Problems 7-9 in Table 8 (the three likelihood of sampling outcomes problems).

Table 8

Problems 7-9 (Likelihood of Sampling Outcomes): Statistical Tests

		Target article		**Replication**			
Problem	Condition	*p*	Cohen's *h* [95% CI]	*p*	Cohen's *h* [95% CI]	BF_{10} (BF_{01})	Replication summary
7 ("Babies")	More extreme	.064	-0.30 [-0.58, -0.03]	.755	-0.06 [-0.20, 0.09]	0.29 (3.50)	No signal–inconsistent, smaller
	Less extreme	.082	-0.30 [-0.60, -0.01]	.332	-0.04 [-0.11, 0.18]	1.80 (0.56)	No signal–inconsistent, smaller
8 ("Investigator")	More extreme	.207	0.20 [-0.08, 0.48]	.013	0.16 [0.02, 0.30]	15 (0.067)	Signal–consistent
	Less extreme	.092	-0.28 [-0.20, 0.37]	.998	-0.23 [-0.37, -0.09]	0.75 (1.30)	No signal–consistent
9 ("Disease")	More extreme	.646	0.09 [-0.20, 0.37]	1.00	-0.30 [-0.44, -0.15]	37690 (0.000)	No signal–inconsistent, opposite
	Less extreme	.646	-0.09 [-0.37, 0.19]	< .001	0.32 [0.18, 0.47]	691 (0.001)	Signal–inconsistent, opposite

Note. One-proportion z-test (one-tailed in the replication). BF = Bayes factor, quantifying evidence for the alternative (BF_{10}) and the null (BF_{01}). Two-tailed *p*-values for the target article and one-tailed *p*-values in the replication. "Smaller" means that the effect is closer to zero.

We ran a series of one-proportion z-tests (one-tailed) for each scenario (babies, investigator, disease) and for each condition ("more extreme" vs. "less extreme") that compared participants' responses against the expected proportion by chance. Following the preregistration, we set the expected proportion at 33.33% (⅓; as per the preregistration).

As these three problems tested a null hypothesis, we also used equivalence testing and Bayesian analysis. We computed Bayes Factor using the BayesFactor R package to quantify evidence in favor of the null hypothesis over the alternative hypothesis (Lee & Wagenmakers, 2005).

Most results were in line with the target article's findings, apart from the "More extreme" condition in Problem 8 and the "Less extreme condition" in Problem 9. In addition, the Bayes Factors were indicative of strong evidence for the alternative hypothesis over the null in both conditions of Problem 9.

Next, we set the expected proportion at 50% (not pre-registered). This is a less conservative test but is arguably more in line with the target article. KT tested whether there was a "significant preference for the correct answer", which we on closer reading interpreted as whether the proportion of correct answers was higher than 50%. Although KT also reported that "About the same" was the modal answer, this is not a statistical test. Note also that according to Teigen (2022) "To test the difference between participants choosing (a) and "equally likely" makes no sense as no meaningful null hypothesis can be formed." (p. 193). With this 50% as the expected proportion, the results are consistent with the target article. That is, the number of participants choosing the correct answer did not exceed 50% in any of the problems.

To further quantify the null in Problems 7-9 (Likelihood of Sampling Outcomes), we examined whether the confidence intervals of each effect in the replication contained the smallest effect size of interest (SESOI). As per the preregistration, we specified the SESOI by halving the smallest effect size in the previous problems in the target article (Problems 1, 2, and 4). Problem 4 in the target article had the smallest effect (Cohen's *h* = 0.39, 95% CI = 0.12, 0.67). Halving this effect size resulted in a Cohen's *h* of 0.20 (95% CI = 0.06, 0.33). We interpreted effects below the lower confidence interval of the SESOI as practically equivalent to zero. Only the effect in the less extreme condition in Problem 7 was lower than the lower confidence interval of the SESOI, suggesting that for the remaining effects, we cannot conclude the absence of an effect.

Next, we conducted an exploratory equivalence test with a less conservative and more common approach. Specifically, we used Simonsohn's (2015) small-telescope approach and defined the SESOI as the effect size the target article had 33% power to detect. This was not preregistered. With this approach, the SESOI was *h* = 0.16. All but two effects had confidence intervals that included 0.16, suggesting that, overall, we cannot conclude the absence of an effect.

Problems 10 and 11 (Posterior Probabilities)

We summarized the descriptives for Problems 10 and 11 in Table 9. We summarized the comparison of the statistical details between the target article's and replication's findings for Problems 10 and 11 (the two posterior probabilities problems) in Table 10.

Table 9

Problems 10 and 11 (Posterior Probabilities Problems): Subjective Probability Estimates

	Target article		**Replication**		
	n	***M***	***n***	***M***	***SD***
Binomial problem: (format: Initial proportion in decks, sample proportion)					
2:3, 18:14	56	58	38	69.74	12.96
2:3, 4:2	56	68	36	71.14	14.74
2:3, 8:4	56	70	39	76.38	13.79
2:3, 40:20	56	70	38	85.76	14.81
2:3, 5:1	56	85	39	76.59	16.64
5:6, 18:14	56	60	36	74.81	16.12
5:6, 4:2	56	70	38	68.47	10.70
5:6, 8:4	56	70	38	75.32	14.27
5:6, 40:20	56	70	39	68.92	11.87
5:6, 5:1	56	83	39	69.28	14.73
Non-binomial problem					
type (i)	115	88.89	378	65.39	26.43
type (ii)	115	71.43	378	56.78	26.19

Note. Subjective Probability Estimates are expressed as percentages. *n* for the binomial problem in the original is the average number of participants in that condition. KT reported that the number of participants for each of the ten conditions in this problem ranged from 37 to 79, with an average of 56.

Table 10

Problems 10 and 11 (Posterior Probabilities): Comparison of target article and replication

		Target article			Replication		
		t	***p***	**Cohen's *d* [95% CI]**	***t***	***p***	**Cohen's *d* [95% CI]**
Binomial problem							
Initial proportion in the decks	Comparison of different sample proportion						
2:1	5:1 vs 4:2	/	<.01	/	2.452	.015	0.56 [0.10, 1.01]
	5:1 vs 8:4	/	<.01	/	2.328	.020	0.53 [0.08, 0.98]
	5:1 vs 40:20	/	<.01	/	-0.064	.949	-0.01 [-0.46, 0.43]
	18:14 vs 4:2	/	<.01	/	0.389	.698	0.09 [-0.36, 0.54]
	18:14 vs 8:4	/	<.01	/	0.252	.801	0.06 [-0.39, 0.50]
	18:14 vs 40:20	/	<.01	/	-2.124	.034	-0.48 [-0.94, -0.03]
5:1	5:1 vs 4:2	/	<.01	/	3.217	.001	0.74 [0.27, 1.20]
	5:1 vs 8:4	/	<.01	/	5.108	<.001	1.16 [0.68, 1.65]
	5:1 vs 40:20	/	<.01	/	3.328	<.001	0.77 [0.30, 1.25]
	18:14 vs 4:2	/	<.01	/	-1.269	.205	-0.29 [-0.75, 0.16]
	18:14 vs 8:4	/	<.01	/	0.568	.499	0.13 [-0.32, 0.58]
	18:14 vs 40:20	/	<.01	/	-1.10	.273	-0.26 [-0.72, 0.20]
Non-binomial problem		/	< .01	/	5.90	<.001	0.30 [0.20 0.41]

Note. For the binomial problem, median tests were conducted in the target article, whereas one-way ANOVA with pairwise comparisons was conducted in the replication. For the non-binomial problem, a median test was conducted in the target article, whereas a paired sample *t*-test was conducted in the replication.

We conducted a one-way ANOVA and Tukey post hoc tests on target comparisons. Recall that KT hypothesized that people would rely on the sample proportion as this is the most representative feature. Specifically, they hypothesized that for both pairs of population proportions (5/6 and 1/6 vs. 2/3 and 1/3), participants' posterior estimates in the 5:1 sample proportion condition would be larger than in the 4:2, 8:4, and 40:20 conditions, which again would be larger than in the 18:14 conditions. This is non-normative: for example, the 40:20 sample provides much stronger evidence than the 5:1 sample. For the conditions with the initial proportion of 2:1 in the deck, we found that the posterior probabilities stated by the participants in conditions 5:1 had no difference from the ones in 4:2, 8:4, and 40:20.

Next, the posterior probabilities stated by the participants in conditions 18:14 also had no difference with the ones in 4:2, 8:4 and 40:20. For the conditions with the initial proportion of 5:1 in the deck, we found that the posterior probabilities stated by the participants in conditions 5:1 were larger than the ones in 4:2, 8:4, and 40:20. The posterior probabilities stated by the participants in condition 18:14 were not different from the ones in 4:2, 8:4 and 40:20 conditions.

The target article found that estimated posterior probabilities in condition 5:1 were larger than those in 4:2, 8:4, and 40:20 for both sets of initial probabilities. Also, estimated posterior probabilities in conditions 18:14 were smaller than those in 4:2, 8:4, and 40:20 for both sets of initial probabilities. However, in the replication, only the estimated posterior probabilities in condition 5:1 were larger than those in 4:2, 8:4, and 40:20 for the initial probability of 5:1. We did not find evidence for differences in the remaining comparisons. Nevertheless, similar to KT, we found that participants were insensitive to population proportions.

For Problem 11 (posterior probabilities, non-binomial), we conducted a paired-sample *t*-test and found that participants attached greater probability to selecting the male population if the sample consisted of a single person whose height was 5 ft 10 in. (case *(i)*) than if the sample consisted of 6 persons whose average height was 5 ft and 8 in. (case *(ii)*), $t(377) = 5.90$, $p < .001$, $d = 0.33$, 95% CI [0.22, 0.44], which is opposite to the normatively correct answer. Our replication results were very similar to those of the target article.

Extension: Decision style

As an extension to the replication, we examined if the decision styles correlated with the extent of using the representativeness heuristic. We calculated reliance on the representativeness heuristic by taking the ratio of scores in Problems 1.1, 1.2, 2, 4, 7, 8, and 9 to the number of heuristic-scoring problems they completed, ranging from 0 to 1 ($M = 0.75$, $SD = 0.22$, Med = 0.75). In our pre-registration, we omitted Problems 1.1, Problem 1.2, and Problem 11 from the calculation because we did not initially recognize that these problems also scored the representativeness heuristic.

We did not find support for the hypothesis that reliance on the representativeness heuristic correlates with intuitive ($r = 0.03$, $p = .501$, 95% CI = -0.05, 0.10) or rational decision style ($r = 0.03$, $p = .476$, 95% CI = -0.05, 0.10). Neither did it correlate with age ($r = .03$, $p = .388$, 95% CI = -0.04, 0.11), gender ($r = -.02$, $p = .581$, 95% CI = -0.10, 0.05), or education ($r = -.01$, $p = .787$, 95% CI = -0.09, 0.06).

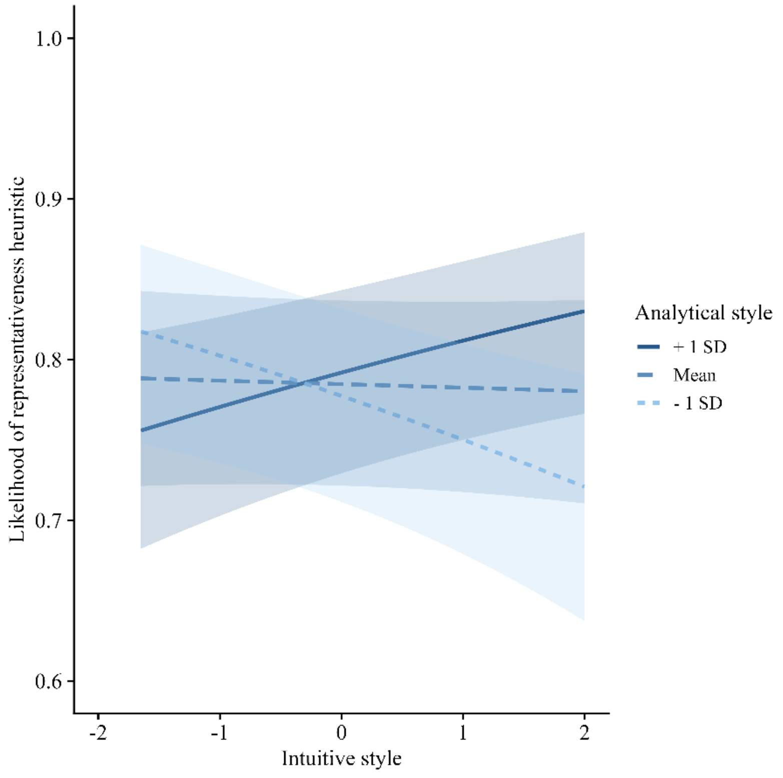
We next examined these associations in a binomial mixed effects model that included problem and subject as random factors, using the *lme4* package in R (Bates et al., 2014). We restructured the data to long format and treated problem as a repeated measure (not preregistered). We did not find an association between the intuitive ($B = 0.02$, $p = .733$, 95% CI = -0.09, 0.13) nor the rational style ($B = 0.14$, $p = .087$, 95% CI = -0.02, 0.31) with the representativeness heuristic.

As an additional exploratory analysis, we examined whether the two styles interactively predicted reliance on the representativeness heuristic. Dual-process theorists suggested that the two styles are conceptually independent and operate interactively (Kahneman, 2002; Norris & Epstein, 2011; Stanovich & West, 2000). Thus, individuals can be grouped into four different categories: high on both styles, low on both styles, high on rationality and low on intuition, and low on rationality and high on intuition (Epstein, 1998; Bakken et al., in press; Hodgkinson & Clarke, 2007; Hodgkinson et al., 2009; Shiloh et al., 2002).

We found a cross-over interaction (*B* = 0.25, *p* = .008, 95% CI = 0.07, 0.43), which we plotted in Figure 3. The interaction plot suggests that those who were high on both dimensions were more prone to using the representativeness heuristic, which is consistent with previous findings (e.g., Shiloh et al., 2002). We will return to these findings in the Discussion.

Figure 3

Interaction Between Intuitive and Rational Styles in Predicting Representativeness Heuristic



Note. Predictors are mean-centered.

Associations and Comparisons Between Problems

One notable strength of the current replication study is that participants completed multiple problems, in contrast to the target article where each problem was presented to a different sample. This setup enabled us to assess the consistency of heuristic responses across problems.

First, we examined the correlations between responses in all of the heuristic-scoring problems (Table 11). We only found evidence for a positive correlation between Problems 7 and 9 and a negative correlation between Problems 4 and 8. These results suggest very poor consistency in participants' responses to the problems.

Next, we explored pairwise comparisons between all problems. We used the *lme4* package (Bates et al., 2014) and ran a logistic mixed effects model with heuristic response (0 = non-heuristic response, 1 = heuristic response) as the dependent variable, problem as the independent variable, and subject as a random factor. The pairwise comparisons using Tukey's test are plotted in Figure 4. Results are given on the log odds ratio scale.

Table 11

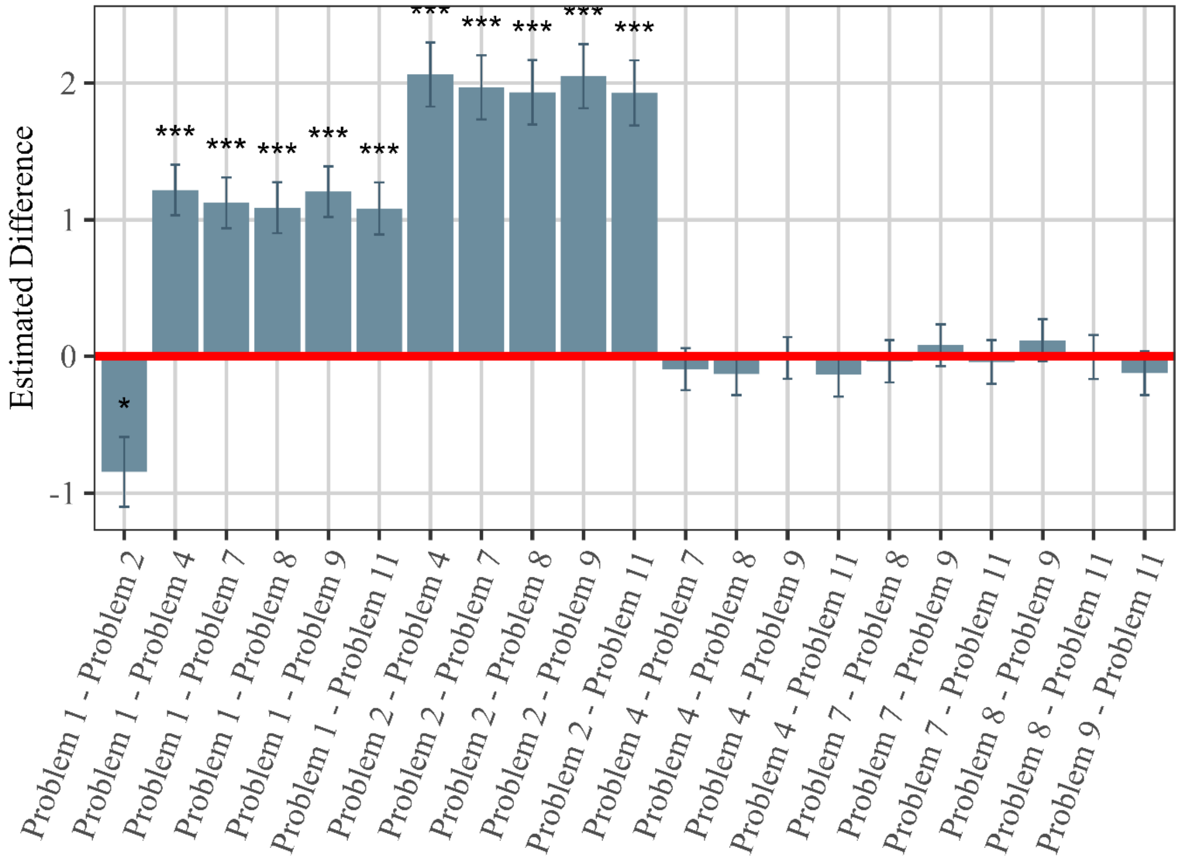
Heuristic Response Problems: Correlations

P#	1.1	1.2	2	4	7	8	9
1.2	.00 [-0.10, 0.10] (377)						
2	-.11 [-0.25, 0.03] (191)	.05 [-0.09, 0.19] (191)					
4	.08 [-0.06, 0.23] (181)	-.07 [-0.22, 0.07] (181)	-.09 [-0.23, 0.04] (199)				
7	-.02 [-0.17, 0.12] (187)	-.01 [-0.15, 0.13] (187)	-.04 [-0.17, 0.09] (221)	-.02 [-0.13, 0.16] (183)			
8	-.13 [-0.27, 0.02] (187)	-.05 [-0.19, 0.10] (187)	.00 [-0.13, 0.13] (221)	-.20** [-0.33, -0.06] (183)	.09 [-0.01, 0.19] (383)		
9	.04 [-0.11, 0.18] (187)	-.02 [-0.17, 0.12] (187)	.10 [-0.03, 0.23] (221)	.01 [-0.14, 0.15] (183)	.22*** [0.13, 0.32] (383)	-.10 [-0.19, 0.01] (383)	
11	-.08 [-0.23, 0.07] (170)	.03 [-0.12, 0.18] (170)	-.06 [-0.21, 0.08] (182)	-.08 [-0.23, 0.07] (171)	.04 [-0.11, 0.20] (160)	.02 [-0.14, 0.17] (160)	-.00 [-0.16, 0.15] (160)

Note. * indicates *p* < .05. ** indicates *p* < .01.

Figure 4

Heuristic Response Problems: Pairwise Comparisons



Note. Problem 1 contains two sub-questions (Problem 1.1 and Problem 1.2), and given that these questions were highly similar and that Problem 1.2 was mainly included as a robustness check, we only included Problem 1.1 here ("Problem 1" in the figure).

Figure 4 indicates that Problem 1 (sample-to-population similarity, birth sequence) differed from almost all of the other problems. Problem 2 (sample-to-population similarity, high-school program) differed slightly from Problem 1, but more from Problems 4-11. We found no support for pairwise comparisons differences among Problems 4-11. A visual inspection of these pairwise comparisons suggest two clusters of problems.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). Hillsdale, N.J: L. Erlbaum Associates.

Hamilton, K., Shih, S., & Mohammed, S. (2016). The development and validation of the rational and intuitive decision styles scale. *Journal of Personality Assessment*. doi:10.1080/00223891.2015.1132426

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*(3), 430-454. doi: https://doi.org/10.1016/0010-0285(72)90016-3

Lakens, D., & Caldwell, A. R. (2021). Simulation-Based Power Analysis for Factorial Analysis of Variance Designs. *Advances in Methods and Practices in Psychological Science, 4*(1), 251524592095150. https://doi.org/10.1177/2515245920951503

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science*, *1*, 389-402.

LeBel, E. P., Vanpaemel, W., Cheung, I., & Campbell, L. (2019). A brief guide to evaluate replications. *Meta-Psychology*, 3.